

Introduction to Statistics - Second Edition

INTRODUCTION TO STATISTICS - SECOND EDITION

An Excel-Based Approach

VALERIE WATTS

Fanshawe College Pressbooks
London Ontario



Introduction to Statistics - Second Edition Copyright © 2025 by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

CONTENTS

Acknowledgements	xi
About this Book	xii
Changes From Previous Version	xvi
Book Navigation	xix

Part I. Sampling and Data

1.1 Definitions of Statistics, Probability, and Key Terms	3
1.2 Types of Data and Levels of Measurement	13
1.3 Sampling and Sampling Techniques	20
1.4 Experimental Design and Ethics	29

Part II. Descriptive Statistics

2.1 Frequency Distributions and Histograms	43
2.2 Measures of Central Tendency	65
2.3 Skewness and the Mean, Median, and Mode	81
2.4 Measures of Position	90
2.5 Measures of Variability	113

Part III. Probability

3.1 The Terminology of Probability	135
3.2 Contingency Tables	145

3.3 The Complement Rule	155
3.4 The Addition Rule	161
3.5 Conditional Probability	174
3.6 Joint Probabilities	192

Part IV. Discrete Probability Distributions

4.1 Random Variables	215
4.2 Probability Distribution of a Discrete Random Variable	223
4.3 Expected Value and Standard Deviation for a Discrete Probability Distribution	234
4.4 The Binomial Distribution	254
4.5 The Poisson Distribution	273

Part V. Continuous Probability Distributions and the Normal Distribution

5.1 Probability Distribution of a Continuous Random Variable	289
5.2 The Normal Distribution	295
5.3 The Standard Normal Distribution	306
5.4 Calculating Probabilities for a Normal Distribution	321

Part VI. The Central Limit Theorem and Sampling Distributions

6.1 Sampling Distribution of the Sample Mean	343
6.2 Sampling Distribution of the Sample Proportion	359

Part VII. Confidence Intervals for Single Population Parameters

7.1 Introduction to Confidence Intervals	381
7.2 Confidence Intervals for a Single Population Mean with Known Population Standard Deviation	384
7.3 Confidence Intervals for a Single Population Mean with Unknown Population Standard Deviation	407
7.4 Confidence Intervals for a Population Proportion	425
7.5 Calculating the Sample Size for a Confidence Interval	441

Part VIII. Hypothesis Tests for Single Population Parameters

8.1 Null and Alternative Hypotheses	457
8.2 The Hypothesis Test Process	465
8.3 Outcomes and the Type I and Type II Errors	479
8.4 Hypothesis Tests for a Population Mean with Known Population Standard Deviation	488
8.5 Hypothesis Tests for a Population Mean with Unknown Population Standard Deviation	505
8.6 Hypothesis Tests for a Population Proportion	522

Part IX. Statistical Inference for Two Populations

9.1 Statistical Inference for Two Population Means with Known Population Standard Deviations	549
9.2 Statistical Inference for Two Population Means with Unknown Population Standard Deviations	571
9.3 Statistical Inference for Matched Samples	600
9.4 Statistical Inference for Two Population Proportions	631

Part X. Statistical Inferences Using the Chi-Square Distribution

10.1 The Chi Square Distribution	657
10.2 Statistical Inference for a Single Population Variance	665
10.3 The Goodness-of-Fit Test	685
10.4 The Test of Independence	710

Part XI. Statistical Inference Using the F-Distribution

11.1 The F-Distribution	735
11.2 Statistical Inference for Two Population Variances	742
11.3 One-Way ANOVA and Hypothesis Tests for Three or More Population Means	769

Part XII. Simple Linear Regression

12.1 Linear Equations	799
12.2 Scatter Diagrams	805
12.3 Correlation	823
12.4 The Regression Equation	840
12.5 Coefficient of Determination	859
12.6 Standard Error of the Estimate	870

Part XIII. Multiple Regression

13.1 Multiple Regression	885
13.2 Standard Error of the Estimate	905
13.3 Coefficient of Multiple Determination	918
13.4 Testing the Significance of the Overall Model	934

13.5 Testing the Regression Coefficients	950
13.6 Multicollinearity	973

Part XIV. Time Series Analysis

14.1 Time Series Patterns	977
14.2 Measures of Forecast Accuracy	995
14.3 Smoothing Models	1019
14.4 Seasonal Indices	1059
14.5 Regression Models	1070

Part XV. Statistical Quality Control

15.1 Control Charts	1101
15.2 Control Charts for Variables	1114
15.3 Control Charts for Attributes	1145
References	1167
Versioning History	1184

ACKNOWLEDGEMENTS

This open textbook has been developed by Dr. Valerie Watts in partnership with the OER Design Studio and the Library Learning Commons at Fanshawe College in London, Ontario.

This work is part of the FanshaweOpen learning initiative and is made available through a Creative Commons Attribution-ShareAlike 4.0 International License unless otherwise noted.



This book is an adaptation of Introduction to Statistics Copyright © 2022 by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

See the Changes from the Previous Version for a list of updates.

Collaborators

This project was a collaboration between the author and the team in the OER Design Studio at Fanshawe. The following staff and students were involved in the creation of this project:

- Catherine Steeves, *Quality Assurance*
- Jason Benoit, *Quality Assurance*
- Shauna Roch, *Project Lead*
- Wilson Poulter, *Copyright*

ABOUT THIS BOOK

Introduction to Statistics: An Excel-Based Approach introduces students to the concepts and applications of statistics, with a focus on using Excel to perform statistical calculations. The book is written at an introductory level and designed for students in fields other than mathematics or engineering who require a fundamental understanding of statistics. The text emphasizes understanding and application of statistical tools over theory, but some knowledge of algebra is required.

Although the text focuses on concepts and applications, every effort has been made to provide both information essential to understanding a topic and sound methodological development. Generally accepted terminology and notation for each topic is used throughout without becoming overly focused on technical details.

In place of manual calculations and the use of probability distribution tables, the text utilizes Excel in the application of statistical analysis. Because of the prevalence and use of Excel in a wide range of fields, it is important for students to understand how to leverage Excel's statistical capabilities. Throughout the text, information is provided on the appropriate Excel function required for a calculation and the solutions to examples illustrate how to use the corresponding Excel function to solve problems.

The text is organized into fifteen chapters and then divided into subchapters by concept or topic. The chapters are as follows:

Chapter 1: Sampling and Data	Chapter 1 covers key definitions, terms, and terminology used in statistics, as well as exploring different sampling methods, types of data, and levels of measurement.
Chapter 2: Descriptive Statistics	Chapter 2 examines the different descriptive statistics, both graphical and numerical, required to organize, summarize, and describe data.
Chapter 3: Probability	Chapter 3 introduces the concept of probability, probability terminology, different approaches to probability, and various probability rules.
Chapter 4: Discrete Random Variables	Chapter 4 explores discrete random variables and their probability distributions, the mean and standard deviation of a discrete random probability distribution, and examines the binomial and Poisson distributions.
Chapter 5: Continuous Random Variables and the Normal Distribution	Chapter 5 covers continuous random variables, focusing on the normal distribution and probability problems associated with the normal distribution.
Chapter 6: The Central Limit Theorem and Sampling Distributions	Chapter 6 examines the sampling distributions of the sample mean and the sampling distribution of the sample proportion.
Chapter 7: Confidence Intervals for Single Population Parameters	Chapter 7 explores the construction, use, and interpretation of confidence intervals to estimate a population mean or a population proportion, as well as determining the sample size necessary for the required accuracy of a confidence interval.
Chapter 8: Hypothesis Tests for Single Population Parameters	Chapter 8 introduces the formal hypothesis testing procedure, focusing on conducting and drawing a conclusion from a hypothesis test on a population mean or a population proportion.
Chapter 9: Statistical Inference for Two Populations	Chapter 9 extends confidence intervals and hypothesis testing to the difference between two population means or the difference between two population proportions.
Chapter 10: Statistical Inference Using the χ^2-Distribution	Chapter 10 covers the use of the χ^2 -distribution in statistical inference, including confidence intervals and hypothesis testing for a population variance, the goodness-of-fit test, and the test of independence.
Chapter 11: Statistical Inference Using the F-Distribution	Chapter 11 examines the use of the F -distribution in statistical inference, including confidence intervals and hypothesis testing for the ratio of two population variances and the one-way ANOVA test on the equality of three or more population means.
Chapter 12: Simple Linear Regression and Correlation	Chapter 12 explores the linear relationship between two variables through the simple linear regression model, including methods to assess the validity of the model.
Chapter 13: Multiple Regression	Chapter 13 extends the linear regression model to include more than one independent variable, including methods to assess the validity of the model.

Chapter 14: Time Series Analysis	Chapter 14 introduces the basics of time series analysis, focusing on various time series forecasting models, such as smoothing models, seasonal indices, and regression models, as well as measures to assess the accuracy of a forecast model.
Chapter 15: Statistical Quality Control	Chapter 15 covers statistical process control and its use of control charts to assess if a process is in-control or out-of-control.

For the Student

Each sub-chapter in this text begins with a list of relevant learning objectives. Where appropriate, videos are included to review, enhance, and extend the material covered in the text. At the end of each sub-chapter, a series of exercises, including answers, are provided to check retention and assess understanding.

Accessibility Statement

We are actively committed to increasing the accessibility and usability of the textbooks we produce. Every attempt has been made to make this OER accessible to all learners and is compatible with assistive and adaptive technologies. We have attempted to provide closed captions, alternative text, or multiple formats for on-screen and offline access.

The web version of this resource has been designed to meet Web Content Accessibility Guidelines 2.0, level AA. In addition, it follows all guidelines in Appendix A: Checklist for Accessibility of the *Accessibility Toolkit – 2nd Edition*.

In addition to the web version, additional files are available in a number of file formats including PDF, EPUB (for eReaders), and MOBI (for Kindles).

If you are having problems accessing this resource, please contact us at oyer@fanshawec.ca.

Please include the following information:

- The location of the problem by providing a web address or page description
- A description of the problem
- The computer, software, browser, and any assistive technology you are using that can help us diagnose and solve your issue (e.g., Windows 10, Google Chrome (Version 65.0.3325.181), NVDA screen reader)

Feedback

To provide feedback on this text, please contact **oer@fanshawec.ca**.

CHANGES FROM PREVIOUS VERSION

This book is the second edition of Introduction to Statistics by Dr. Valerie Watts, licensed under Creative Commons Attribution NonCommercial ShareAlike.

The following is a summary of changes between the first edition and the second edition:

Overall	<ul style="list-style-type: none"> • Moved introductory sub-chapters to chapter title page. • Added additional exercises as necessary. • Removed concept reviews from the end of each sub-chapter. • Moved exercises from the end of each chapter to appropriate sub-chapters. • Added answers to all exercises. • Added chapters on time series analysis and statistical quality control.
Chapter 1: Sampling and Data	<ul style="list-style-type: none"> • Moved levels of measurement to sub-chapter on types of data. • Moved frequency distributions to chapter 2. • Created a new sub-chapter for sampling and sampling techniques.
Chapter 2: Descriptive Statistics	<ul style="list-style-type: none"> • Removed frequency polygons. • Moved time series plots to chapter 14. • Changed order of sub-chapters. • Removed content on grouped data calculations.
Chapter 3: Probability	<ul style="list-style-type: none"> • Added additional exercises for the complement rule.
Chapter 4: Discrete Random Variables	<ul style="list-style-type: none"> • Added additional content on random variables.
Chapter 5: Continuous Random Variables and the Normal Distribution	<ul style="list-style-type: none"> • Combined sub-chapters on continuous random variables and continuous probability distributions.

Chapter 14: Time Series Analysis	<ul style="list-style-type: none">• Created this new chapter.
Chapter 15: Statistical Quality Control	<ul style="list-style-type: none">• Created this new chapter.

BOOK NAVIGATION


Recommended Format: Online Webbook

You can access this resource online using a desktop computer or mobile device or download it for free on the main landing page of this resource. Look for the “Download this book” drop-down menu directly below the webbook cover. This resource is available for download in the following formats:

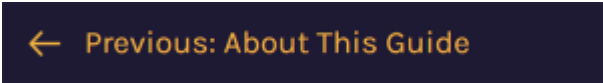
- **PDF.** You can download this book as a PDF to read on a computer (Digital PDF) or print it out (Print PDF). The digital PDF preserves hyperlinks and provides default navigation within the document. In addition, the PDF allows the user to highlight, annotate, and zoom the text.
- **Mobile.** If you want to read this textbook on your phone or tablet, use the EPUB (eReader) or MOBI (Kindle) files. Please refer to your device’s features for additional support when navigating this resource.

Navigating this Webbook

To move to the next page, click on the “Next” button at the bottom right of your screen.

A dark blue rectangular button with the text "Next: 1.1. What is Academic Integrity? →" in orange.

To move to the previous page, click on the “Previous” button at the bottom left of your screen.

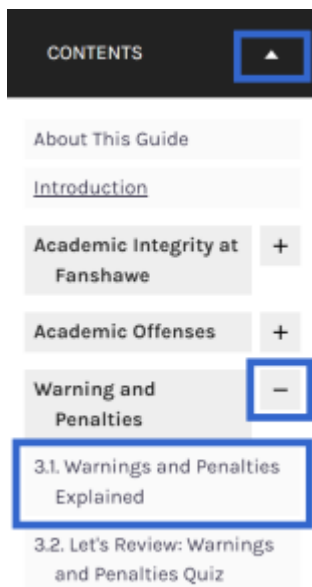
A dark blue rectangular button with the text "← Previous: About This Guide" in orange.

Keyboard arrows can also be used to navigate. *(Note: On smaller screens, the “Previous” and “Next” buttons are stacked at the bottom of the page.)*

To scroll back up to the top of the page, click on the bottom middle of your screen (*Note: this will only appear if the page is long*).



To jump to a specific section or sub-section, click on “Contents” in the top left section of the page. Use the plus sign (+) to expand and the minus sign (-) to collapse the content sections. (*Note: On smaller screens, the “Contents” button is at the top of the page.*)



PART I

SAMPLING AND DATA

You are probably asking yourself the question, “When and where will I use statistics?” If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate, just to mention a few. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or “fact.” Statistical methods can help you make the “best educated guess.”

Because you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what “good” data can be distinguished from “bad.”

CHAPTER OUTLINE

- 1.1 Definitions of Statistics, Probability, and Key Terms
- 1.2 Types of Data and Levels of Measurement
- 1.3 Sampling and Sampling Techniques
- 1.4 Experimental Design and Ethics

“1.1 Introduction to Sampling and Data” from Introduction to Statistics by Valerie Watts is licensed

under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

1.1 DEFINITIONS OF STATISTICS, PROBABILITY, AND KEY TERMS

LEARNING OBJECTIVES

- Recognize and differentiate between key terms used in statistics.

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives. The organization and summation of data is called **descriptive statistics**. Two ways to summarize data are by graphing, such as a histogram or box plot, and by using numbers, such as average and standard deviation. After we have studied probability and probability distributions, we will use formal methods for drawing conclusions from “good” data. These formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data, or inference, is based on good procedures for producing data and thoughtful examination of the data. Although there are numerous mathematical formulas for analyzing data, the goal of statistics is to gain an understanding of the data, and not to simply perform calculations using the formulas. These days, we use computers to perform the calculations. The understanding and interpretation of the data comes from us. If we can thoroughly grasp the basics of statistics, we can be more confident in the decisions you make in life.

Probability

Probability is a mathematical tool used to study randomness and the chance, or likelihood, of an event occurring. For example, if we toss a **fair** coin four times, the outcomes may not necessarily be two heads and two tails. However, if we toss the same coin 4,000 times, the outcomes will be close to

half heads and half tails. The expected theoretical probability of heads in any one toss is 50%. Even though the outcomes from a small number of repetitions are uncertain, there is a regular pattern to the outcomes when there is a large number of repetitions.

The theory of probability began with the study of games of chance, such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether a student will get an A in a particular course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. We might use probability to decide whether or not to buy a lottery ticket. In the study of statistics, we use the power of mathematics through probability calculations to analyze and interpret the data.

Key Terms

In statistics, we generally want to study a population. A **population** is a collection of persons, things, or objects under study. Because populations tend to be very large, it is too expensive and too time-consuming to study the entire population. Instead of studying the population, we study a **sample** taken from the population. The idea of **sampling** is to select a portion, or subset, of the larger population and study that sample to gain information about the population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. Consider the following examples:

- If we wished to compute the overall grade point average at a school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages.
- In federal elections, opinion polls typically sample between 1,000 and 2,000 people. The opinion poll is supposed to represent the views of the people in the entire country.
- Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can actually contains 16 ounces of a carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average grade earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a number that is a property of the population. Because we considered all math classes to be the population, then the average grade earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A variable, notated by capital letters such as X or Y , is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. Numerical variables take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal one student's grade in a math class at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then some examples of Y include Conservative, Liberal, and New Democrat. In this case, Y is a categorical variable. We could do some math with values of X , such as calculate the average grade, but it makes no sense to do math with values of Y . **Data** are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If we take three exams written by a single student in a math class and obtain scores of 86, 75, and 92, we would calculate the student's mean score by adding the three exam scores and dividing by three, in this case a mean of 84.3. If a class has 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

NOTE

The words **mean** and **average** are often used interchangeably. The substitution of one word for the other is common practice. The technical term for mean is “arithmetic mean,” and “average” is technically a centre of location. However, in practice among non-statisticians, “average” is commonly accepted for “arithmetic mean.”

EXAMPLE

Determine the key terms for the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first-year students at the college. Three of those students spent \$150, \$200, and \$225, respectively.

Solution

- The **population** is all first year students attending ABC College this term.
- The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).
- The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term (the population mean).
- The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample (the sample mean).
- The **variable** could be the amount of money spent (excluding books) by one first year student. Let X be the amount of money spent (excluding books) by one first year student attending ABC College.
- The **data** are the dollar amounts spent by the first year students. Examples of the data are \$150, \$200, and \$225.

EXAMPLE

Determine the key terms refer for the following study. As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which cars crashed	56 kilometres/hour
Location of crash test dummies	Front seat

Cars with dummies in the front seats were crashed into a wall at a speed of 56 kilometres per hour. We want to know the proportion of dummies in the driver's seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution

- The **population** is all cars containing dummies in the front seat.
- The **sample** is the 75 cars, selected by a simple random sample.
- The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.
- The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.
- The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.
- The **data** are either: yes, had head injury, or no, did not.

EXAMPLE

Determine the key terms for the following study. An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Solution

- The **population** is all medical doctors listed in the professional directory.
- The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

- The **sample** is the 500 doctors selected at random from the professional directory.
- The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.
- The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.
- The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

TRY IT

Determine the key terms for the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent \$65, \$75, and \$95, respectively.

Click to see Solution

- The **population** is all families with children attending Knoll Academy.
- The **sample** is a random selection of 100 families with children attending Knoll Academy.
- The **parameter** is the average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.
- The **statistic** is the average (mean) amount of money spent on school uniforms by families in the sample.
- The **variable** is the amount of money spent by one family. Let X be the amount of money spent on school uniforms by one family with children attending Knoll Academy.
- The **data** are the dollar amounts spent by the families. Examples of the data are \$65, \$75, and \$95.

TRY IT

Determine the key terms for the following study. A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

- | | | |
|--------------------|-------------------|------------------|
| 1. Population ____ | 3. Parameter ____ | 5. Variable ____ |
| 2. Statistic ____ | 4. Sample ____ | 6. Data ____ |
- all students who attended the college last year.
 - the cumulative GPA of one student who graduated from the college last year.
 - 3.65, 2.80, 1.50, 3.90.
 - a randomly selected group of students who graduated from the college last year.
 - the average cumulative GPA of all students who graduated from the college last year.
 - all students who graduated from the college last year.
 - the average cumulative GPA of students in the study who graduated from the college last year.

Click to see Solution

- | | | |
|------|------|------|
| 1. f | 3. e | 5. b |
| 2. g | 4. d | 6. c |

Exercises

- Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. A researcher selects 40 patients with AIDS from the start of treatment until their deaths. Identify the population, the sample, the parameter, the sample and the variable for this study.

Click to see Answer

- Population: all AIDS patients
- Sample: the 40 AIDS patients in the study.
- Parameter: the average length of time (in months) AIDS patients live after treatment.
- Sample: the average length of time (in months) patients in the study live after treatment.
- Variable: the length of time an individual AIDS patient lives after treatment.

2. For each of the following eight exercises, identify the key terms: the population, the sample, the parameter, the statistic, and the variable.
- a. A fitness centre is interested in the mean amount of time a client exercises in the centre each week.
 - b. Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.
 - c. A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.
 - d. Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.
 - e. A politician is interested in the proportion of voters in his district who think he is doing a good job.
 - f. A marriage counsellor is interested in the proportion of clients she counsels who stay married.
 - g. Political pollsters may be interested in the proportion of people who will vote for a particular cause.
 - h. A marketing company is interested in the proportion of people who will buy a particular product.

Click to see Answer

- a.
 - Population: all clients at the fitness centre.
 - Sample: a group or subset of the clients at the fitness centre.
 - Parameter: the mean amount of time all clients at the fitness centre exercise each week.
 - Statistic: the mean amount of time the clients in the sample exercise each week.
 - Variable: the amount of time an individual client at the fitness centre exercises each week.
- b.
 - Population: all children who take ski or snowboard lessons.
 - Sample: a group or subset of children who take ski or snowboard lessons.

- Parameter: the mean age of all children who take ski or snowboard lessons.
 - Statistic: the mean age of the children in the sample.
 - Variable: the age of an individual child who takes ski or snowboard lessons.
- c.
- Population: all of the cardiologist's patients who have had heart attacks.
 - Sample: a group or subset of cardiologist's patients who have had heart attacks.
 - Parameter: the mean recovery time of all patients who have had heart attacks.
 - Statistic: the mean recovery time of the patients in the sample.
 - Variable: the recovery time of an individual patient who has had a heart attack.
- d.
- Population: all clients at the insurance company.
 - Sample: a group or subset of clients at the insurance company.
 - Parameter: the mean health cost of all clients at the insurance company.
 - Statistic: the mean health cost of the clients in the sample.
 - Variable: the health cost of an individual client at the insurance company.
- e.
- Population: all voters in the politician's district.
 - Sample: a group or subset of voters in the politician's district.
 - Parameter: the proportion of all voters in the district who think the politician is doing a good job.
 - Statistic: the proportion of the voters in the sample who think the politician is doing a good job.
 - Variable: the number of voters who think the politician is doing a good job.
- f.
- Population: all clients of the counsellor.
 - Sample: a group or subset of the counsellor's clients.
 - Parameter: the proportion of all clients who stayed married.
 - Statistic: the proportion of clients in the sample who stayed married.
 - Variable: the number of clients who stayed married.
- g.
- Population: all voters.
 - Sample: a group or subset of voters.
 - Parameter: the proportion of all voters who vote for the cause.
 - Statistic: the proportion of voters in the sample who vote for the cause.
 - Variable: the number of voters who vote for the cause.
- h.
- Population: all consumers.

- Sample: a group or subset of consumers.
- Parameter: the proportion of all consumers who purchase the product.
- Statistic: the proportion of consumers in the sample who purchase the product.
- Variable: the number of consumers who purchase the product.

3. A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.
- a. What is the population she is interested in?
 - b. Consider the following: X = number of days a Lake Tahoe Community College math student is absent. In this case, X is an example of what?
 - c. The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of what?

Click to see Answer

- a. All math students at Lake Tahoe Community College.
- b. A variable.
- c. A statistics.

“1.2 Definitions of Statistics, Probability, and Key Terms” and “1.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

1.2 TYPES OF DATA AND LEVELS OF MEASUREMENT

LEARNING OBJECTIVES

- Identify data as qualitative or quantitative.
- Classify data by level of measurement.

Types of Data

Data may come from a population or from a sample. Generally, small letters like x or y are used to represent data values. Most data can be put into one of two categories: qualitative or quantitative.

Qualitative data are the result of categorizing or describing attributes of a population. Qualitative data are also called **categorical data**. Hair colour, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair colour might be black, dark brown, light brown, blonde, grey, or red. Blood type might be AB+, O-, or B+.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. The amount of money, pulse rate, weight, the number of people living in your town, and the number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair colour or blood type.

All data that are the result of counting are called **quantitative discrete data**. These data take on

only certain numerical values. For example, the number of phone calls received in a day could be zero, one, two, or three.

All data that are the result of measuring are **quantitative continuous data**, assuming that we can measure accurately. Measuring angles in radians might result in such numbers as $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π , $\frac{3\pi}{4}$ and so on. For example, the number of books in a backpack is discrete data, and the weight of the backpack is continuous data.

Levels of Measurement

In addition to being classified as quantitative or qualitative, data is classified into four levels of measurement. The way a set of data is measured is called its **level of measurement**. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be applied to every set of data. Qualitative data has a level of either nominal scale or ordinal scale. Quantitative data has a level of either interval scale or ratio scale.

Data that is measured using a **nominal scale** is data that can be placed into categories. Colours, names, labels, favourite foods, and yes/no survey responses are examples of nominal level data. Nominal scale data are **not** ordered, which means the categories of the data are not ordered. For example, trying to “order” people according to their favourite food does not make any sense. Putting pizza first and sushi second is not meaningful. Smartphone companies are another example of nominal scale data. Some examples are Sony, Motorola, Nokia, Samsung, and Apple. This is just a list of different brand names, and there is no agreed upon order for the categories. Some people may prefer Apple, but that is a matter of opinion. Because nominal data consists of categories, nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data in that the data can be placed into categories, but there is a big difference. The categories of ordinal scale data can be ordered or ranked. An example of ordinal scale data is a list of the top five national parks in the country because the parks can be ranked from one to five. Another example of using the ordinal scale is a cruise survey, where the responses to questions about the cruise are “excellent,” “good,” “satisfactory,” and “unsatisfactory.” These responses are ordered from the most desired response to the least desired. In ordinal scale data, the differences between two pieces of data cannot be measured or calculated. Similar to nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using an **interval scale** is similar to ordinal level data because it has a definite ordering. However, the differences between interval scale data can be measured or

calculated, but the data does not have a starting point. Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements (Celsius and Fahrenheit), 40° is equal to 100° minus 60° . The differences in temperature can be measured and make sense. But there is no starting point to the temperature scales because 0° is not the absolute lowest temperature. Temperatures like -10°F and -15°C exist, and are colder than 0° . Interval level data can be used in calculations, but ratios do not make sense and cannot be done. For example, 80°C is not four times as hot as 20°C (nor is 80°F four times as hot as 20°F). So there is no meaning to the ratio of 80 to 20 (or four to one) in either temperature scale. In general, ratios have no meaning in interval scale data.

Data that is measured using the **ratio scale** takes care of the ratio problem and gives us the most information. Ratio scale data is like interval scale data, but it has a starting point to the scale (a 0 point), and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The data can be put in order from lowest to highest: 20, 68, 80, 92. The differences between the data have meaning: 92 minus 68 is 24. Ratios can be calculated: 80 is four times 20. The smallest possible score is 0.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=28#oembed-1>

Video: “Nominal, ordinal, interval and ratio data: How to Remember the differences” by NurseKillam [11:03] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

EXAMPLE

The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books (three, four, two, and one) are quantitative discrete data. The level of measurement is ratio.

EXAMPLE

The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, and 4.3. Weights are quantitative continuous data. The level of measurement is ratio.

EXAMPLE

Classify each of the following by type of data (qualitative or quantitative) and level of measurement (nominal, ordinal, interval, ratio). For quantitative data, classify as discrete or continuous.

1. High school soccer players are classified by their athletic ability: Superior, Average, and Above average. **Solution:** qualitative, ordinal.
2. Baking temperatures for various main dishes. **Solution:** quantitative, discrete, interval.

3. The colours of crayons in a crayon box. **Solution:** qualitative, nominal.
4. The heights of 21-65 year-old women. **Solution:** quantitative, continuous, ratio.
5. Common letter grades: A, B, C, D, F. **Solution:** qualitative, ordinal.

TRY IT

Classify each of the following by type of data (qualitative or quantitative) and level of measurement (nominal, ordinal, interval, ratio). For quantitative data, classify as discrete or continuous.

1. Number of machines in a gym.
2. Social insurance numbers.
3. Incomes measured in dollars.
4. A satisfaction survey of a social website by number: 1 = very satisfied, 2 = somewhat satisfied, 3 = not satisfied.
5. Political outlook: extreme left, left-of-centre, right-of-centre, extreme right.
6. Time of day on an analog watch.
7. The distance in miles to the closest grocery store.
8. The dates 1066, 1492, 1644, 1947, and 1944.
9. Areas of lawns in square meters.

Click to see Solution

1. Quantitative, discrete, ratio.
2. Qualitative, nominal.
3. Quantitative, discrete, ratio.
4. Qualitative, ordinal.
5. Qualitative, nominal.
6. Quantitative, discrete, interval.
7. Quantitative, continuous, ratio.

8. Qualitative, ordinal.
9. Quantitative, continuous, ratio.

Exercises

1. Classify each of the following as qualitative or quantitative. For quantitative data, classify as discrete or continuous. Identify the level of measurement for each data.
 - a. Number of times per week.
 - b. Size of automobile (compact, midsize, large).
 - c. Age.
 - d. Weight of package shipped.
 - e. Temperature.
 - f. Satisfaction rating (good, fair, poor).
 - g. Shoe size.
 - h. Attendance at home games.
 - i. Country of origin.
 - j. Number on team uniform.
 - k. Number of tickets sold to a concert.
 - l. Percent of body fat.
 - m. Favourite baseball team.
 - n. Time in line to buy groceries.
 - o. Number of students enrolled at Evergreen Valley College.
 - p. Most-watched television show.
 - q. Brand of toothpaste.
 - r. Distance to the closest movie theatre.
 - s. Number of competing computer spreadsheet software packages.

Click to see Answer

- a. Quantitative, discrete, ratio
- b. Qualitative, nominal.
- c. Quantitative, discrete, ratio.
- d. Quantitative, discrete, ratio.
- e. Quantitative, continuous, interval.

- f. Qualitative, ordinal.
- g. Quantitative, discrete, interval.
- h. Quantitative, discrete, ratio.
- i. Qualitative, nominal.
- j. Qualitative, nominal.
- k. Quantitative, discrete, ratio.
- l. Quantitative, continuous, ratio.
- m. Qualitative, nominal.
- n. Quantitative, continuous, ratio.
- o. Quantitative, discrete, ratio.
- p. Qualitative, ordinal.
- q. Qualitative, nominal.
- r. Quantitative, continuous, ratio.
- s. Quantitative, discrete, ratio.

“1.3 Sampling and Data“, “1.4 Frequency, Frequency Tables, and Levels of Measurement” and “1.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

1.3 SAMPLING AND SAMPLING TECHNIQUES

LEARNING OBJECTIVES

- Apply various types of sampling methods to data collection.

Gathering information about an entire population often costs too much, is too time consuming, or is virtually impossible. Instead, we use a sample of the population. In order to get accurate conclusions about the population from the sample, a sample should have the **same characteristics as the population it represents**. Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods.

There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. In simple random sampling, any group of n individuals is equally likely to be chosen as any other group of n individuals. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members, not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in the following table.

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Ludquist	23	Rowell
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cunningham	16	Okimoto	26	Stratcher
06	Fontecha	17	Patel	27	Tallai
07	Hong	18	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai
09	Jiao	20	Reyes	30	Wood
10	Khan				

Lisa can use a computer to generate random numbers. Suppose the computer generates the following numbers:

14 05 04

The number 14 corresponds to Macierz, the number 05 corresponds to Cunningham, and the number 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Macierz, Cunningham, and Cuarismo.

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.

To choose a **stratified sample**, divide the population into groups called **strata**, and then take a **proportionate** number from each stratum. For example, a college's student population can be stratified (grouped) by department, and then a proportionate simple random sample is chosen from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups), and then randomly select some of the clusters. All the members from the selected clusters are in the cluster sample. For example, divide a college's faculty by department, so the departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers form the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n th piece of data from a listing of the population. For example, a phone book contains 20,000 residential listings, from which 400 names must be selected. Number the population from 1 to 20,000, and then use a simple random sample to pick a number that represents the first name in the sample. Then, choose every fiftieth name thereafter until a total of 400 names are selected. Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is **non-random** is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. Such a sample is not random because only those customers in the store on that particular day have the opportunity to be in the sample. The results of convenience sampling may be very good in some cases and highly biased (favour certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased because they may favour a certain group. It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However, for practical reasons, in most populations, simple random sampling is done **without replacement**, where a member of the population may only be chosen once or not at all. Surveys are typically done without replacement. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Consequently, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

In a college population of 10,000 people, suppose we want to randomly pick a sample of 1,000 for a survey. For any particular sample of 1,000, if we are sampling **with replacement** (where the person is replaced before picking the next person):

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a **different** second person for this sample is 999 out of 10,000 (0.0999),

and the chance of picking the same person again is 1 out of 10,000 (very low).

If we are sampling **without replacement** (where the person is not replaced before picking the next person):

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a **different** second person is 999 out of 9,999 (0.0999), and the chance of picking the same person again is 0.

Comparing the fractions $\frac{999}{10,000}$ and $\frac{999}{9,999}$ to four decimal places, these numbers are equivalent. So we can see that the chance of selecting a small sample from a large population is basically the same, whether or not the sampling is done with replacement.

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is 10, and we are sampling **with replacement for any particular sample**, then the chance of picking the first person is 10 out of 25, and the chance of picking a **different** second person is 9 out of 25 (we replace the first person). If we sample without replacement, then the chance of picking the first person is still 10 out of 25, but the chance of picking the second person (who is different) is 9 out of 24. Comparing the fractions $\frac{9}{25} = 0.36$ and $\frac{9}{24} = 0.3750$, these numbers are not equivalent.

When we analyze data, it is important to be aware of **sampling errors** and **non-sampling errors**. The actual process of sampling causes sampling error, which is the difference between the actual population parameter and the corresponding sample statistic. In reality, a sample will never be exactly representative of the population, so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error. Factors not related to the sampling process cause **non-sampling errors**. For example, a defective counting device can cause a non-sampling error.

In statistics, a **sampling bias** is created when a sample is collected from a population, and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=375#oembed-1>

Video: “Statistics: Sources of Bias” by Mathispower4u [4:44] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A study was done to determine the age, number of times per week, and the duration (amount of time) residents use a local park. The first house in the neighbourhood around the park was selected randomly, and then the resident of every eighth house in the neighbourhood around the park was interviewed.
 - a. What sampling method was used?
 - b. “Duration (amount of time)” is what type of data?
 - c. The colours of the houses around the park are what kind of data?

Click to see Answer

- a. Systematic.
 - b. Quantitative, continuous.
 - c. Qualitative.
2. For the following four exercises, determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).
 - a. A group of test subjects is divided into twelve groups; then, four of the groups are chosen at random.
 - b. A market researcher polls every tenth person who walks into a store.
 - c. The first 50 people who walk into a sporting event are polled on their television preferences.
 - d. A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen.

Click to see Answer

- a. Cluster.

- b. Systematic.
- c. Convenience.
- d. Simple Random.

3. Identify the sampling method used in each of the following situations.

- a. A woman in the airport is handing out questionnaires to travellers, asking them to evaluate the airport's service. She does not ask travellers who are hurrying through the airport with their hands full of luggage but instead asks all travellers who are sitting near gates and not taking naps while they wait.
- b. A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.
- c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.
- d. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.
- e. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.
- f. The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. What type of sampling was used?

Click to see Answer

- a. Convenience.
- b. Cluster.
- c. Stratified.
- d. Systematic.
- e. Simple Random.
- f. Stratified.

4. Suppose you want to determine the mean number of students per statistics class in your

college. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

Click to see Answer

(Answers will vary.) You could use a cluster sampling method. Each statistics class is a cluster. Randomly select some of the clusters/classes and record the number of students in each of the selected classes.

5. Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

Click to see Answer

(Answers will vary.) You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then, stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

6. List some practical difficulties involved in getting accurate results from a telephone survey.

Click to see Answer

(Answers will vary.) Many people will simply hang up. Many people will not pick up the phone if they do not recognize the caller ID. If they do respond to the surveys, you cannot be sure who is responding. Many people will not be called at all because they only have cell phones and phone lists generally only contain landline numbers.

7. List some practical difficulties involved in getting accurate results from a mailed survey.

Click to see Answer

(Answers will vary.) Many people will not respond to mail surveys. If they do respond to the surveys, you cannot be sure who is responding. In addition, mailing lists can be incomplete.

8. Airline companies are interested in the consistency of the number of babies on each flight so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the

result of that study.

- a. Using complete sentences, list three things wrong with the way the survey was conducted.
- b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

Click to see Answer

- a. The survey was conducted using six similar flights. The survey would not be a true representation of the entire population of air travellers.
- b. Conduct the survey during different times of the year. Conduct the survey using flights to and from various locations.
Conduct the survey on different days of the week.

9. In advance of the 1936 Presidential Election, a magazine titled Literary Digest released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent postcards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists.

Approximately 2,300,000 people returned the postcards.

- a. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
- b. What effect does the low response rate have on the reliability of the sample?
- c. Are these problems examples of sampling error or nonsampling error?
- d. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. His researchers used a method they called “quota sampling” to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this section?

Click to see Answer

- a. This 1936 poll was conducted during the Great Depression. Most of the population at that time could not afford magazines, cars, phones or club memberships, and so would not receive the postcards. These people had no chance of being included in the poll.
- b. The low response rate means the sample may not be representative of the entire population.
- c. Non-sampling.
- d. Stratified.

“1.3 Sampling and Data” and “1.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

1.4 EXPERIMENTAL DESIGN AND ETHICS

LEARNING OBJECTIVES

- Describe different aspects of experimental design.
- Apply ethical behaviour in statistical analysis.

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? Questions like these are answered using randomized experiments. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes a change in another, we call the first variable the **explanatory variable**. The affected variable is called the **response variable**. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable are called **treatments**. An **experimental unit** is a single object or individual to be measured.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=32#oembed-1>

Video: “Observational Studies and Experiments” by ProfessorMcComb [3:06] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Suppose we want to investigate the effectiveness of vitamin E in preventing disease. We recruit a group of subjects and ask them if they regularly take vitamin E. We notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? No, it does not. There are many differences between the two groups compared, in addition to vitamin E consumption. People who take vitamin E often take other steps to improve their health, such as exercise, diet, other vitamin supplements, or choosing not to smoke. Any one of these factors could be influencing a person's health. As described, this study does not prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called **lurking variables**. In order to prove that the explanatory variable is the cause of a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design their experiment in such a way that there is only one difference between the groups being compared: the planned treatments. This is accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point, the only difference between groups is the one imposed by the researcher. Therefore, different outcomes measured in the response variable must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment—a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if someone is participating in a study and they know that they are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, they do not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

EXAMPLE

Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks.

Identify the following values for this study: population, sample, experimental units, explanatory variable, response variable, and treatments.

Solution

- The **population** is men aged 50 to 84.
- The **sample** is the 400 men who participated.
- The **experimental units** are the individual men in the study.
- The **explanatory variable** is the oral medication.
- The **treatments** are aspirin and a placebo.
- The **response variable** is whether a subject had a heart attack.

EXAMPLE

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first

three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

1. Describe the explanatory and response variables in this study.
2. What are the treatments?
3. Identify any lurking variables that could interfere with this study.
4. Is it possible to use blinding in this study?

Solution

1. The explanatory variable is scent, and the response variable is the time it takes to complete the maze.
2. There are two treatments: a floral-scented mask and an unscented mask.
3. All subjects experienced both treatments. The order of treatments was randomly assigned, so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.
4. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. However, researchers timing the mazes can be blinded. The researcher who is observing a subject will not know which mask is being worn.

EXAMPLE

A researcher wants to study the effects of birth order on personality. Explain why this study could not be conducted as a randomized experiment. What is the main problem in a study that cannot be designed as a randomized experiment?

Solution

The explanatory variable is birth order. We cannot randomly assign a person's birth order. Random assignment eliminates the impact of lurking variables. When we cannot assign subjects to treatment groups at random, there will be differences between the groups other than the explanatory variable.

TRY IT

A researcher wants to study the effects of texting on driving performance. The researcher designs a study to test the response time of drivers while texting and while driving only and measures how many seconds it takes for a driver to respond when a leading car hits the brakes?

1. Describe the explanatory and response variables in the study.
2. What are the treatments?
3. What should the researcher consider when selecting participants?
4. The researcher considers dividing participants randomly into two groups: one to drive without distraction and one to text and drive simultaneously. Is this a good idea? Why or why not?
5. Identify any lurking variables that could interfere with this study.
6. How can blinding be used in this study?

Click to see Solution

1. The explanatory variable is texting, and the response variable is time to hit the brakes.
2. There are two treatments: driving while texting and driving only.
3. Participants should be experienced drivers.
4. This is not a good idea. The purpose of the study is to test the effect of texting on driving. To draw a meaningful conclusion, each driver must perform the test for both treatments: driving while texting and driving only.
5. If each participant experiences both treatments and the treatments are randomly assigned, there are no lurking variables.
6. The participants cannot be blinded because they will know if they are texting while driving or driving only. The researcher timing the drivers could potentially be blinded.

Ethics

The widespread misuse and misrepresentation of statistical information often gives the field a bad name. Some say that “numbers don’t lie,” but the people who use numbers to support their claims often do.

A recent investigation of famous social psychologist Diederik Stapel has led to the retraction of his articles from some of the world's top journals, including the *Journal of Experimental Social Psychology*, *Social Psychology*, *Basic and Applied Social Psychology*, *British Journal of Social Psychology*, and the magazine *Science*. Diederik Stapel is a former professor at Tilburg University in the Netherlands. Recently, an extensive investigation involving three universities where Stapel worked concluded that the psychologist is guilty of fraud on a colossal scale. Falsified data taints over 55 papers he authored and 10 Ph.D. dissertations that he supervised.

The committee investigating Stapel concluded that he is guilty of several practices, including:

- creating datasets, which largely confirmed the prior expectations;
- altering data in existing datasets;
- changing measuring instruments without reporting the change and
- misrepresenting the number of experimental subjects.

Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not as easy to spot.

Researchers have a responsibility to verify that proper methods are being followed. Many of Stapel's co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

Many types of statistical fraud are difficult to spot. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They do not want to take the chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis.

Professional organizations, like the American Statistical Association, clearly define expectations for researchers. There are even laws in the federal code about the use of research data.

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects. Most countries have federal regulations that protect participants in research studies. When a university or other research institution engages in research, it must ensure the safety of all human subjects. For this reason, research institutions establish oversight committees, commonly known as **Institutional Review Boards (IRB)**. All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy.

These ideas may seem fundamental, but they can be very difficult to verify in practice. Is removing a participant's name from the data record sufficient to protect privacy? Perhaps the person's identity could be discovered from the data that remains. What happens if the study does not proceed as planned and risks arise that were not anticipated? When is informed consent really necessary? Suppose your doctor wants a blood sample to check your cholesterol level. Once the sample has been tested, you expect the lab to dispose of the remaining blood. At that point, the blood becomes biological waste. Does a researcher have the right to take it for use in a study?

It is important that students of statistics take time to consider the ethical questions that arise in statistical studies. How prevalent is fraud in statistical studies? You might be surprised—and disappointed. There is a website dedicated to cataloguing retractions of study articles that have been proven fraudulent. A quick glance will show that the misuse of statistics is a bigger problem than most people realize.

Vigilance against fraud requires knowledge. Learning the basic theory of statistics empowers users and observers to analyze statistical studies critically.

EXAMPLE

Describe the unethical behaviour in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A researcher is collecting data in a community.

1. She selects a block where she is comfortable walking because she knows many of the people living on the street.
2. No one seems to be home at four houses on her route. She does not record the addresses and does not return at a later time to try to find residents at home.

3. She skips four houses on her route because she is running late for an appointment. When she gets home, she fills in the forms by selecting random answers from other residents in the neighbourhood.

Solution

1. By selecting a convenient sample, the researcher is intentionally selecting a sample that could be biased. Claiming that this sample represents the community is misleading. The researcher needs to select areas in the community at random.
2. Intentionally omitting relevant data will create bias in the sample. Suppose the researcher is gathering information about jobs and child care. By ignoring people who are not home, she may be missing data from working families that are relevant to her study. She needs to make every effort to interview all members of the target sample.
3. It is never acceptable to fake data. Even though the responses she uses are “real” responses provided by other participants, the duplication is fraudulent and can create bias in the data. She needs to work diligently to interview everyone on her route.

TRY IT

Describe the unethical behaviour, if any, in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A study is commissioned to determine the favourite brand of fruit juice among teens in California.

1. The survey is commissioned by the seller of a popular brand of apple juice.
2. There are only two types of juice included in the study: apple juice and cranberry juice.
3. Researchers allow participants to see the brand of juice as samples are poured for a taste test.
4. Twenty-five percent of participants prefer Brand X, 33% prefer Brand Y, and 42% have no preference between the two brands. Brand X references the study in a commercial, saying, “Most teens like Brand X as much as or more than Brand Y.”

Click to see Solution

1. The seller of the apple juice has a vested interest in the outcome of the study. Their participation in the study may result in misleading or biased data.
2. The study is intentionally omitting relevant data. What if a teen's favourite fruit juice is orange? The study misses data that is relevant to the study.
3. Because participants can see the brand, familiarity or preference for one brand over another may influence participants to pick that brand, regardless of which juice flavour they prefer. This introduces bias into the sample.
4. The company is making a fraudulent claim about the data in the commercial.

Exercises

1. Discuss potential violations of the rule requiring informed consent.
 - a. Inmates in a correctional facility are offered good behaviour credit in return for participation in a study.
 - b. A research study is designed to investigate a new children's allergy medication.
 - c. Participants in a study are told that the new medication being tested is highly promising, but they are not told that only a small portion of participants will receive the new medication. Others will receive placebo treatments and traditional treatments.

Click to see Answer

- a. Inmates may not feel comfortable refusing participation or may feel obligated to take advantage of the promised benefits. They may not feel truly free to refuse participation.
 - b. Parents can provide consent on behalf of their children, but children are not competent to provide consent for themselves.
 - c. All risks and benefits must be clearly outlined. Study participants must be informed of relevant aspects of the study in order to give appropriate consent.
2. How does sleep deprivation affect your ability to drive? A recent study measured the effects on 19 professional drivers. Each driver participated in two experimental sessions: one after normal sleep and one after 27 hours of total sleep deprivation. The treatments were assigned in random order. In each session, performance was measured on a variety of tasks, including a driving simulation. Use key terms from this chapter to describe the design of this

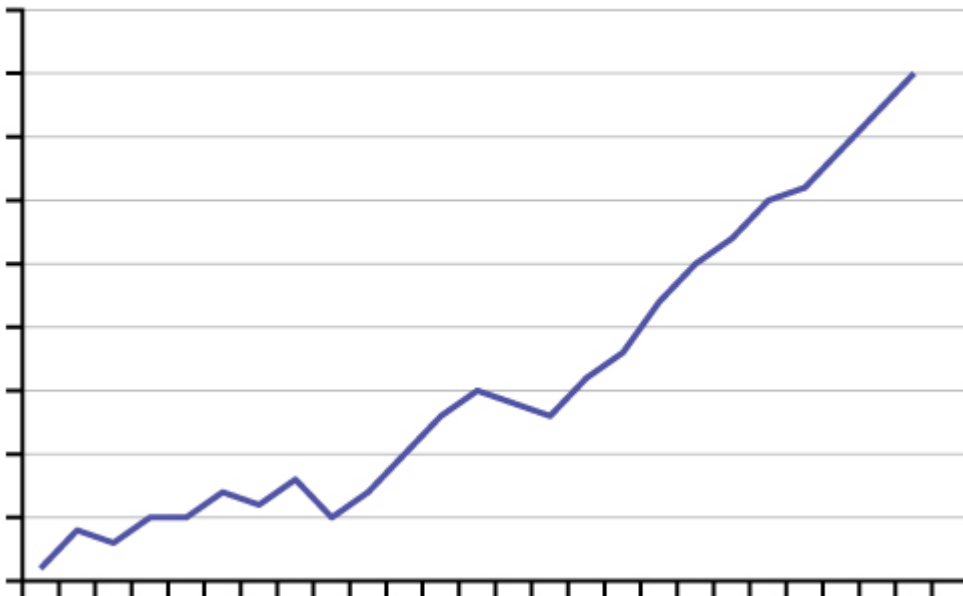
experiment.

Click to see Answer

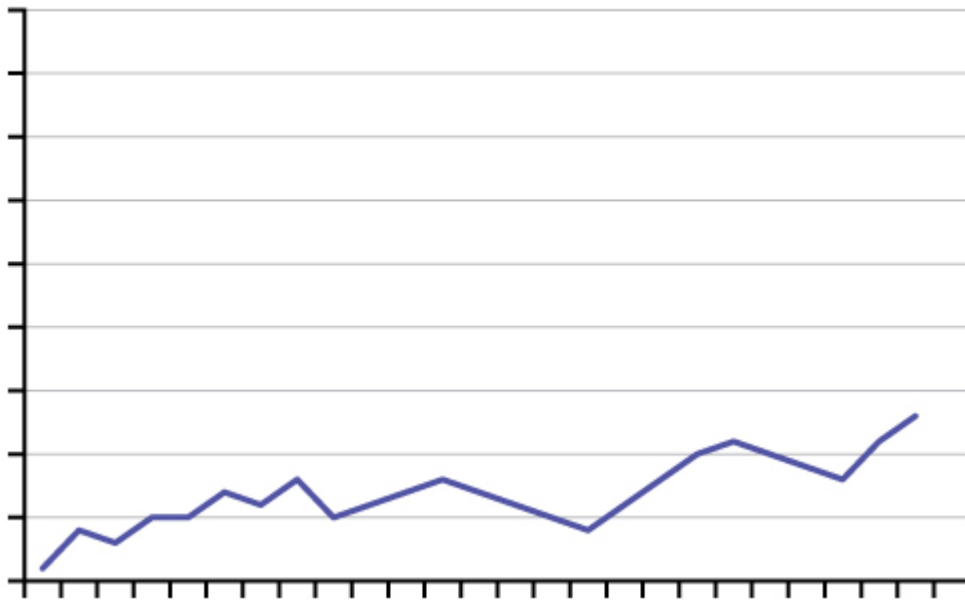
- Explanatory variable: the amount of sleep.
- Response variable: performance measured in assigned tasks.
- Treatments: normal sleep and 27 hours of total sleep deprivation.
- Experimental units: 19 professional drivers.
- Lurking variables: none – all drivers participated in both treatments.
- Random assignment: treatments were assigned in random order; this eliminated the effect of any “learning” that may take place during the first experimental session.
- Control/Placebo: completing the experimental session under normal sleep conditions.
- Blinding: researchers evaluating subjects’ performance must not know which treatment is being applied at the time⁸⁹. You cannot assume that the numbers of complaints reflect the quality of the airlines. The airlines shown with the greatest number of complaints are the ones with the most passengers. You must consider the appropriateness of methods for presenting data; in this case, displaying totals is misleading.

3. An advertisement for Acme Investments displays the two graphs in the figure below to show the value of Acme’s product in comparison with the Other Guy’s product. Describe the potentially misleading visual effect of these comparison graphs. How can this be corrected?

Acme Investments



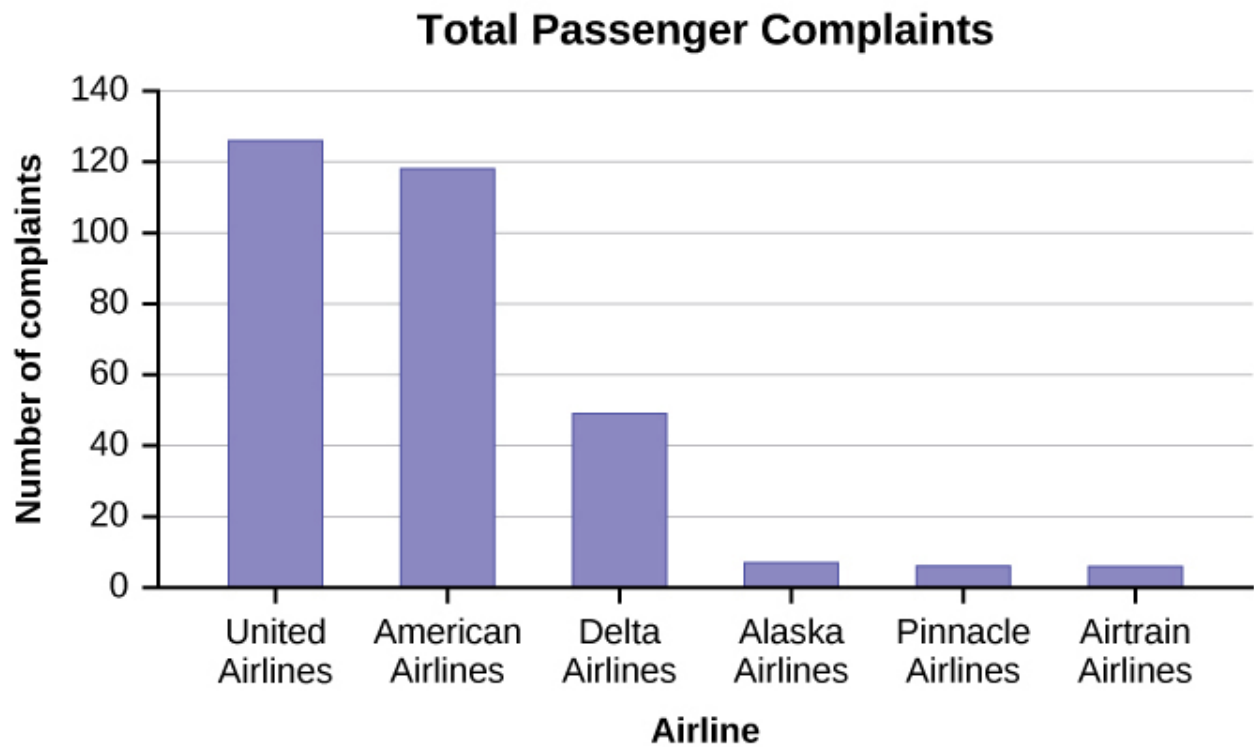
Other Guy's Investments



Click to see Answer

There is no scale on either graph, making it impossible to tell which one produced the best return on investments. The results are shown in separate graphs, making it difficult to compare the results directly. To fix this, plot both graphs on the same set of axes and include a scale on the axes.

4. The graph in the figure below shows the number of complaints for six different airlines as reported to the US Department of Transportation in February 2013. Alaska, Pinnacle, and Airtran Airlines have far fewer complaints reported than American, Delta, and United. Can we conclude that American, Delta, and United are the worst airline carriers since they have the most complaints?

**Click to see Answer**

Because American, United and Delta are the largest carriers, with the largest number of total passengers, and so will have the largest number of complaints. A more accurate graph would be to show the proportion of complaints out of the total number of passengers.

“1.5 Experimental Design and Ethics” and “1.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

PART II

DESCRIPTIVE STATISTICS

Once we have collected data, what do we do with it? Data can be described and presented in many different formats. For example, suppose we are interested in buying a house in a particular area. We may have no clue about the house prices, so we might ask a real estate agent to give us a sample data set of prices. Looking at all the prices in the sample is often overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that we can summarize and describe data. The real estate agent might also provide us with a graph of the data. In this chapter, we will study numerical and graphical ways to describe and display your data. This area of statistics is called **descriptive statistics**. We will learn how to calculate and, more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps us learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data, and then use more formal tools to analyze the data.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the pie chart, and the box plot. In this chapter, we will briefly look at bar graphs (or histograms).

CHAPTER OUTLINE

- 2.1 Frequency Distributions and Histograms
- 2.2 Measures of Central Tendency
- 2.3 Skewness and the Mean, Median, and Mode
- 2.4 Measures of Position
- 2.5 Measures of Variability

“2.1 Introduction to Descriptive Statistics” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

2.1 FREQUENCY DISTRIBUTIONS AND HISTOGRAMS

LEARNING OBJECTIVES

- Create and interpret frequency distribution tables.
- Display, analyze, and interpret data presented in a histogram.

Once we have a set of data, we need to organize it so that we can analyze how frequently each datum occurs in the set. However, when calculating the frequency, we may need to round our answers so that they are as precise as possible.

Frequency Distributions

A **frequency distribution** or **frequency table** is a summary table of data that shows the number of observations that fall into each of several, non-overlapping classes.

Consider the following data. Twenty students were asked how many hours they worked per day. Their responses, in hours, are recorded in the table below.

5	6	3	3	2	4	7	5	2	3
5	6	5	4	4	3	5	2	5	3

Using each data value as the class, the following table lists the different data values in ascending order and their frequencies.

CLASS	FREQUENCY
2	3
3	5
4	3
5	6
6	2
7	1

A **frequency** is the number of observations in the data that fall into each class. According to the table, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column is 20, which is the total number of students included in the sample.

A **relative frequency** is the ratio, fraction, or proportion of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample—in this case 20. Relative frequencies can be written as fractions, percents, or decimals. The sum of the values in the relative frequency column is 1 or 100\%.

CLASS	FREQUENCY	RELATIVE FREQUENCY
2	3	$\displaystyle\frac{3}{20}\times 100\%=15\%$
3	5	$\displaystyle\frac{5}{20}\times 100\%=25\%$
4	3	$\displaystyle\frac{3}{20}\times 100\%=15\%$
5	6	$\displaystyle\frac{6}{20}\times 100\%=30\%$
6	2	$\displaystyle\frac{2}{20}\times 100\%=10\%$
7	1	$\displaystyle\frac{1}{20}\times 100\%=5\%$

Cumulative frequency is the accumulation of the previous frequencies. To find the cumulative frequencies, add all of the previous frequencies to the frequency for the current row, as shown in the table below. The last entry of the cumulative frequency column is the number of observations in the data.

CLASS	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE FREQUENCY
2	3	15\%	3
3	5	25\%	$3 + 5 = 8$
4	3	15\%	$8 + 3 = 11$
5	6	30\%	$11 + 6 = 17$
6	2	10\%	$17 + 2 = 19$
7	1	5\%	$19 + 1 = 20$

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in the table below. The last entry of the cumulative relative frequency column is 1 or 100\%, indicating that 100\% of the data has been accumulated.

CLASS	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	15\%	3	15\%
3	5	25\%	8	$15\% + 25\% = 40\%$
4	3	15\%	11	$40\% + 15\% = 55\%$
5	6	30\%	17	$55\% + 30\% = 85\%$
6	2	10\%	19	$85\% + 10\% = 95\%$
7	1	5\%	20	$95\% + 5\% = 100\%$

NOTES

1. Because of rounding of the relative frequencies, the relative frequency column may not always sum to 1 or 100\%, and the last entry in the cumulative relative frequency column may not be 1 or 100\%. However, they each should be close to 1 or 100\%. If all of the decimals are kept in the calculations, the relative frequency column will sum to 1 or 100\%

, and the last cumulative relative frequency will be **1** or 100\%.

2. A simple way to round off answers is to carry the final answer to one more decimal place than was present in the original data. Round off only the final answer. Do not round off any intermediate results, if possible. If it becomes necessary to round off intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores four, six, and nine is **6.3**, rounded off to the nearest tenth because the data are whole numbers. Most answers will be rounded off in this manner.

EXAMPLE

The following data are the number of books bought by **50** part-time college students at ABC College. Construct a frequency distribution table for this data using six classes.

1	1	1	1	1	1	1	1	1	1
1	2	2	2	2	2	2	2	2	2
2	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	4	4	4
4	4	4	5	5	5	5	5	6	6

Solution

CLASS	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
1	11	22\%	11	22\%
2	10	20\%	21	42\%
3	16	32\%	37	74\%
4	6	12\%	43	86\%
5	5	10\%	48	96\%
6	2	4\%	50	100\%

Frequency Distributions with Classes

Instead of using a single data value as the class, frequency distributions are often constructed using non-overlapping classes and the frequency is the number of observations from the data that fall into each class. Classes are most often used when constructing a frequency distribution for quantitative data. When setting up the classes, it is important to make sure that the classes do not overlap and that every data value falls into one of the classes. In other words, each data value must fall into one and only one of the classes.

There are three steps to creating the classes for a frequency distribution:

1. **Determine the number of non-overlapping classes.** The number of classes is somewhat subjective and is often determined by the researcher. Although there are no rules about how many classes a frequency distribution should have, general guidelines recommend using between five and twenty classes. For a smaller data set, use a smaller number of classes. For a larger data set, use a larger number of classes. When deciding on the number of classes, the goal is to have enough classes to capture the trends and patterns in the data but not so many classes that some classes contain only a few data values.
2. **Determine the width of the classes.** As with the number of classes, there are no rules about the class width. However, general guidelines recommend that the classes all have the same width because this reduces the user may misinterpret the data. An approximate class

width may be found using the following formula:

$$\text{Approximate Class Width} = \frac{\text{Maximum Data Value} - \text{Minimum Data Value}}{\text{Number of Classes}}$$

Round the number obtained from this formula to a more convenient number, for example, a whole number. For example, if the result of this formula is 14.2, a convenient class width is 15.

3. **Determine the class limits.** Using the number of classes and the class width, construct the class limits. When constructing the classes, remember that every data value must fall into one and only one class. The lower class limit is the smallest possible data value that falls into that class. The upper class limit is the largest possible data value that falls into that class.

NOTES

1. The classes must be non-overlapping and exhaustive. In other words, each data value in the data set must fall into one and only class.
2. To prevent any misinterpretation of the data, the classes should all have the same width.
3. The number of classes and the class width are subjective. It is up to the researcher to determine the number of classes and the class width. Different people may construct completely different, but equally valid, frequency distributions.

EXAMPLE

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. Construct a frequency distribution table with eight classes for this data.

60	64	64.5	66	66.5	67	67.5	69	70	71
60.5	64	64.5	66	66.5	67	67.5	69	70	71
61	64	64.5	66.5	66.5	67	67.5	69	70	72
61	64	66	66.5	67	67	68	69	70	72
61.5	64	66	66.5	67	67	68	69	70	72
63.5	64.5	66	66.5	67	67.5	69	69.5	70	72.5
63.5	64.5	66	66.5	67	67.5	69	69.5	70.5	72.5
63.5	64.5	66	66.5	67	67.5	69	69.5	70.5	73
64	64.5	66	66.5	67	67.5	69	69.5	70.5	73.5
64	64.5	66	66.5	67	67.5	69	69.5	71	74

Solution

1. There will be 8 classes (given in the instructions).
2. Calculate the class width. The minimum data value is 60, and the maximum data value is 74.

$$\begin{aligned}\text{Approximate Class Width} &= \frac{74 - 60}{8} \\ &= 1.75\end{aligned}$$

Rounding this value to 2, each class will have a width of 2.

3. Determine the classes. The first class can start at 60 and go up to, but not include, 62. The second class will start at 62 and go up to, but not include, 64. The third class will start at 64 and go up to, but not include, 66. This process continues until the eighth class that starts at 74 and goes up to, but does not include, 76. (Remember, the classes must be non-overlapping, so we do not want numbers like 62 to potentially fall into both the first and second classes. By writing the classes this way (62 up to but not including 64), we ensure that a number like 62 will only fall into the second class.)

CLASS	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
60 to less than 62	5	5\%	5	5\%
62 to less than 64	3	3\%	8	8\%
64 to less than 66	15	15\%	23	23\%
66 to less than 68	40	40\%	63	63\%
68 to less than 70	17	17\%	80	80\%
70 to less than 72	12	12\%	92	92\%
72 to less than 74	7	7\%	99	99\%
74 to less than 76	1	1\%	100	100\%

TRY IT

The following data are the shoe sizes of 50 male students. Construct a frequency distribution table using four classes.

9	9	9.5	9.5	10	10	10	10	10	10
10.5	10.5	10.5	10.5	10.5	10.5	10.5	10.5	11	11
11	11	11	11	11	11	11	11	11	11
11	11.5	11.5	11.5	11.5	11.5	11.5	11.5	12	12
12	12	12	12	12	12.5	12.5	12.5	12.5	14

Click to see Solution

$$\begin{aligned}
 \text{Approximate Class Width} &= \frac{14 - 9}{4} \\
 &= 1.25 \\
 &\rightarrow \text{Round to 2}
 \end{aligned}$$

CLASS	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
8 to less than 10	4	8\%	4	8\%
10 to less than 12	34	68\%	38	76\%
12 to less than 14	11	22\%	49	98\%
14 to less than 16	1	2\%	50	100\%

Histograms

For most of the work we do in this book, we will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A **histogram** is a visual display of a frequency distribution table. It consists of contiguous, vertical boxes with both a horizontal axis and a vertical axis. The horizontal axis is labelled with the classes or categories from the frequency distribution table. The vertical axis is labelled either **frequency** or

relative frequency (or percent frequency or probability). The graph will have the same shape with either label on the vertical axis but the scale on the vertical axis will be different. The histogram gives us the shape of the data, the center of the data, and the spread of the data.

Recall that the frequency is the number of times an observation falls into that particular class, and the relative frequency is the frequency for the class divided by the total number of data values in the sample. For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then the frequency of the 90% to 100% class is 3 and the relative frequency is $\displaystyle\frac{3}{40}\times 100\%=7.5\%$. So, 7.5% of the students received between 90% and 100%.

EXAMPLE

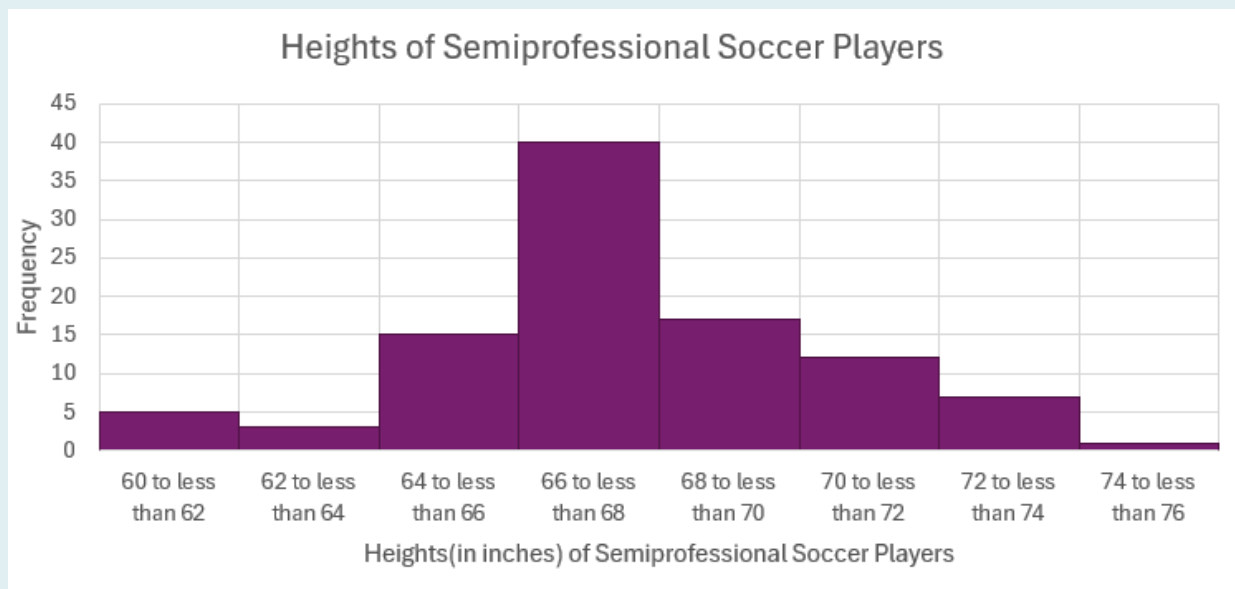
The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. Construct a histogram for this data using eight classes.

60	64	64.5	66	66.5	67	67.5	69	70	71
60.5	64	64.5	66	66.5	67	67.5	69	70	71
61	64	64.5	66.5	66.5	67	67.5	69	70	72
61	64	66	66.5	67	67	68	69	70	72
61.5	64	66	66.5	67	67	68	69	70	72
63.5	64.5	66	66.5	67	67.5	69	69.5	70	72.5
63.5	64.5	66	66.5	67	67.5	69	69.5	70.5	72.5
63.5	64.5	66	66.5	67	67.5	69	69.5	70.5	73
64	64.5	66	66.5	67	67.5	69	69.5	70.5	73.5
64	64.5	66	66.5	67	67.5	69	69.5	71	74

Solution

In a previous example, we constructed the following frequency distribution table. Here, only the frequency column is shown because this is the column that will go on the vertical axis of the histogram.

CLASS	FREQUENCY
60 to less than 62	5
62 to less than 64	3
64 to less than 66	15
66 to less than 68	40
68 to less than 70	17
70 to less than 72	12
72 to less than 74	7
74 to less than 76	1



EXAMPLE

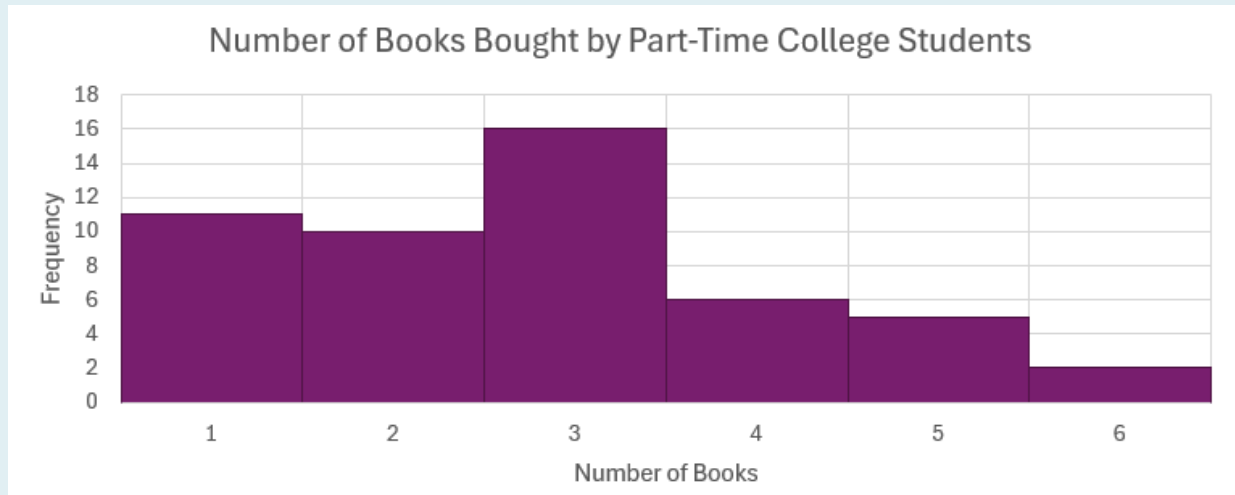
The following data are the number of books bought by 50 part-time college students at ABC College. Construct a histogram for this data using six classes.

1	1	1	1	1	1	1	1	1	1
1	2	2	2	2	2	2	2	2	2
2	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	4	4	4
4	4	4	5	5	5	5	5	6	6

Solution

In a previous example, we constructed the following frequency distribution table. Here, only the frequency column is shown because this is the column that will go on the vertical axis of the histogram.

CLASS	FREQUENCY
1	11
2	10
3	16
4	6
5	5
6	2



CREATING A FREQUENCY DISTRIBUTION AND HISTOGRAM IN EXCEL

In order to create a frequency distribution and its corresponding histogram in Excel, we need to use the Analysis ToolPak. Follow these instructions to add the Analysis ToolPak.

1. Enter your data into a worksheet.
2. Determine the classes for the frequency distribution. Using these classes, create a **Bin** column that contains the **upper limit** for each class.
3. Go to the **Data** tab and click on **Data Analysis**. If you do not see **Data Analysis** in the **Data** tab, you will need to install the Analysis ToolPak.
4. In the **Data Analysis** window, select **Histogram**. Click **OK**.
5. In the **Input** range, enter the cell range for the data.
6. In the **Bin** range, enter the cell range for the **Bin** column.
7. Select the location where you want the output to appear.
8. Select **Chart Output** to produce the corresponding histogram for the frequency distribution.
9. Click **OK**.

This website provides additional information on using Excel to create a frequency distribution.

NOTE

The histogram produced by Excel uses the frequency column from the frequency table on the vertical axis, not the relative frequency column.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=386#oembed-1>

Video: “Frequency Tables, Bar Charts, Pie Charts, Histograms, Grouped & Ungrouped Data Distributions” by Joshua Emmanuel [8:41] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=386#oembed-2>

Video: “How to Construct a Histogram in Excel using built-in Data Analysis” by Joshua Emmanuel [1:59] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	60\%	
2	15		
3			

- Fill in the blanks in the table
- What percent of students take exactly two courses?
- What percent of students take one or two courses?

Click to see Answer

a.

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	60%	60%
2	15	30%	90%
3	5	10%	100%

- 30%
- 90%

2. Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown below:

# Flossing per Week	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0	27	45\%		
1	18			
3				93.33\%
6	3	5\%		
7	1	1.67\%		

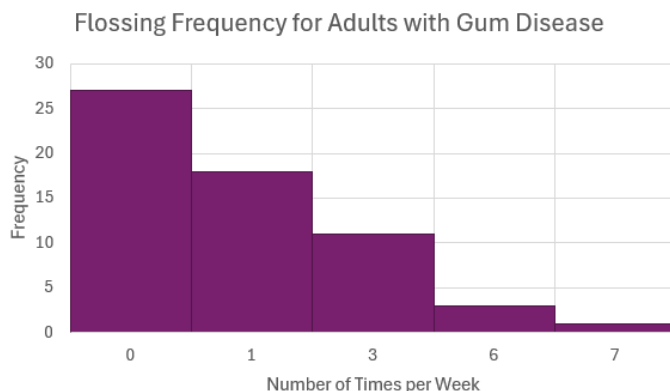
- Fill in the blanks in the table.
- How many adults flossed more than three times per week?
- What percent of adults flossed six times per week?
- What percent of adults flossed at most three times per week?
- Construct the histogram for this frequency distribution table.

Click to see Answer

a.

# Flossings per Week	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0	27	45\%	27	45\%
1	18	30\%	45	75\%
3	11	18.33\%	56	93.33\%
6	3	5\%	59	98.33\%
7	1	1.67\%	60	100\%

- 4
- 5\%
- 93.33\%



e.

3. Forbes magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, have a stock price of at least \$5 per share, and have reported annual revenue between \$5 million and \$1 billion. The table below shows the ages of the chief executive officers for the first 60 ranked firms.

Class	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
40 – 44	3			
45 – 49	11			
50 – 54	13			
55 – 59	16			
60 – 64	10			
65 – 69	6			
70 – 74	1			

- Complete the frequency distribution table.
- How many CEOs are between 55 and 64 years of age?
- What percentage of CEOs are 65 years or older?
- What percentage of CEOs are under 50 years of age?
- How many CEOs are younger than 55 years of age?
- Construct the histogram for this frequency distribution.

Click to see Answer

a.

Class	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
40 – 44	3	5\%	3	5\%
45 – 49	11	18.33\%	14	23.33\%
50 – 54	13	21.67\%	27	45\%
55 – 59	16	26.67\%	43	71.67\%
60 – 64	10	16.67\%	53	88.33\%
65 – 69	6	10\%	59	98.33\%
70 – 74	1	1.67\%	60	100\%

- b. 26
- c. 11.67\%
- d. 23.33\%
- e. 27



f.

4. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the length of time in months patients live once starting the treatment. A researcher follows 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected.

16	3	4	11	15	16	17	22	44	37
14	24	25	15	26	27	33	29	35	44
13	21	22	10	12	8	40	32	26	27
31	34	29	17	8	24	18	47	33	34

- a. Construct a frequency distribution for this data using eight classes.
- b. How many patients live between 15 and 20 months?
- c. How many patients live more than 32 months?
- d. What percentage of patients live between 27 and 44 months?
- e. What percentage of patients live less than 15 months?
- f. Construct the histogram for this frequency distribution.

Click to see Answer

The answer will vary depending on the chosen classes.

a.

Class	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
3 – 8	4	10\%	4	10\%
9 – 14	5	12.5\%	9	22.5\%
15 – 20	7	17.5\%	16	40\%
21 – 26	8	20\%	24	60\%
27 – 32	6	15\%	30	75\%
33 – 38	6	15\%	36	90\%
39 – 44	3	7.5\%	39	97.5\%
45 – 50	1	2.5\%	40	100\%

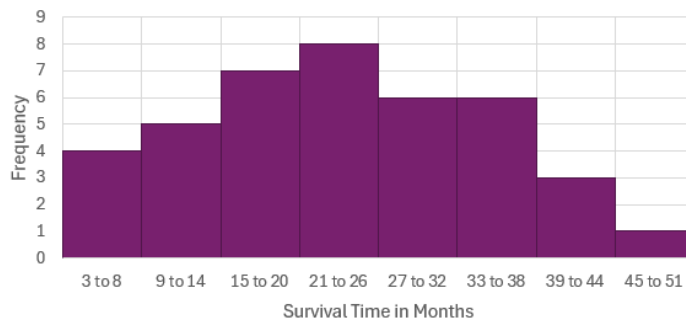
b. 7

c. 10

d. 37.5\%

e. 22.5\%

Length of Time AIDS Patients Live After Starting Treatment



f.

5. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows.

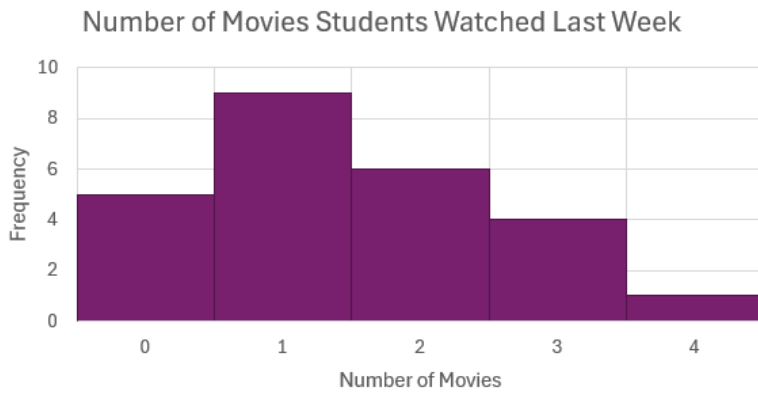
Number of Movies	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0	5			
1	9			
2	6			
3	4			
4	1			

- Complete the frequency distribution table.
- How many students watched less than two movies last week?
- How many students watched between one and three movies last week?
- What percentage of students watched one or more movies last week?
- What percentage of students watched three or four movies last week?
- Construct the histogram for the frequency distribution table.

Click to see Answer

a.	Number of Movies	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
	0	5	20\%	5	20\%
	1	9	36\%	14	56\%
	2	6	24\%	20	80\%
	3	4	16\%	24	96\%
	4	1	4\%	1	100\%

- 14
- 19
- 80\%
- 20\%



f.

6. How much time does it take to travel to work in a particular region? The table below shows the commute time for a sample of workers in the region who are at least 16 years old and do not work at home.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

- Construct the frequency distribution table for this data using six classes.
- Which class contains the most data? What percentage of the data does this represent?
- What percentage of commuters in the sample take the longest to get to work?
- How many commuters in the sample take the least amount of time to get to work?
- Construct the corresponding histogram for the frequency distribution.

Click to see Answer

The answer will vary depending on the chosen classes.

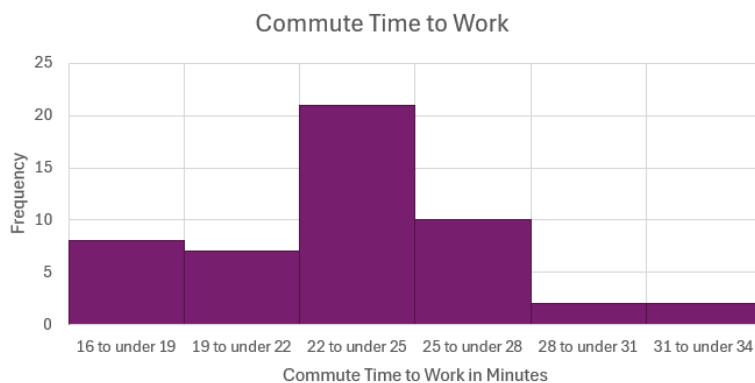
a.

Classes	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
16 to under 19	8	16\%	8	16\%
19 to under 22	7	14\%	15	30\%
22 to under 25	21	42\%	36	72\%
25 to under 28	10	20\%	46	92\%
28 to under 31	2	4\%	48	96\%
31 to under 34	2	4\%	50	100\%

b. 21 to under 25; 42\%

c. 4\%

d. 8



e.

“2.2 Histograms, Frequency Polygons, and Time Series Graphs” and “2.7 Exercices” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

2.2 MEASURES OF CENTRAL TENDENCY

LEARNING OBJECTIVES

- Recognize, describe, calculate, and analyze the measures of the centre of data: mean, median, and mode.

The “centre” of a data set is a way of describing location. The two most widely used measures of the “centre” of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data, and find the number that splits the data into two equal parts so that half of the numbers are below the median and the other half of the numbers are above the median. The median is generally a better measure of the center when there are **extreme values** or **outliers** because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the centre.

NOTE

The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term for mean is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average” is commonly accepted for “arithmetic mean.”

Mean

The **mean** is calculated by adding up all of the values in the data and then dividing the sum by the total number of data values.

The letter used to represent the sample mean is \bar{x} (read x -bar). The Greek letter μ (pronounced “mew”) represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

Consider the sample:

1	1	1	2	2	3	4	4	4	4	4
---	---	---	---	---	---	---	---	---	---	---

$$\bar{x} = \frac{1 + 1 + 1 + 2 + 2 + 3 + 4 + 4 + 4 + 4 + 4}{11} = 2.7$$

CALCULATING THE MEAN IN EXCEL

To find the mean in Excel, use the **average(array)** function.

- For **array**, enter the array or cell range containing the data.

The output from the **average** function is the mean of the entered data.

Visit the Microsoft page for more information about the **average** function.

Median

The **median** is the middle value in an **ordered** set of data. You can quickly find the **location** of the median by using the expression $\frac{n+1}{2}$ where n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then

the median is located in position $\frac{n+1}{2} = \frac{97+1}{2} = 49$ of the ordered list. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5$ and the median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median.

CALCULATING THE MEDIAN IN EXCEL

To find the median in Excel, use the **median(array)** function.

- For **array**, enter the array or cell range containing the data.

The output from the **median** function is the median of the entered data.

Visit the Microsoft page for more information about the **median** function.

EXAMPLE

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3	4	8	8	10	11	12	13	14	15
15	16	16	17	17	18	21	22	22	24
24	25	26	26	27	27	29	29	31	32
33	33	34	34	35	37	40	44	44	47

Calculate the mean and the median.

Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A40.

For the mean:

Function	average
Field 1	A1:A40
Answer	23.575 months

For the median:

Function	median
Field 1	A1:A40
Answer	24 months

TRY IT

The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3	4	5	7	7	7	7	8	8	9
9	10	10	10	10	10	11	12	12	13
14	14	15	15	17	17	18	19	19	19
21	21	22	22	23	24	24	24	24	

Click to see Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A39.

For the mean:

Function	average
Field 1	A1:A39
Answer	13.949 months

For the median:

Function	median
Field 1	A1:A39
Answer	13 months

EXAMPLE

Suppose that in a small town of 50 people, one person earns \$5,000,000 per year and the other 49 each earn \$30,000. Which is the better measure of the “centre”: the mean or the median?

Solution

$$\mu = \frac{5,000,000 + (49 \times 30,000)}{50} = \$129,400$$

$$M = \$30,000$$

The median is a better measure of the “centre” than the mean because 49 of the values are \$30,000

and one is \$5, 000, 000. The \$5, 000, 000 is an outlier. The median of \$30, 000 gives us a better sense of the middle of the data.

TRY IT

In a sample of 60 households, one house is worth \$2, 500, 000. Half of the rest are worth \$280, 000, and all the others are worth \$315, 000. Which is the better measure of the “centre”: the mean or the median?

Click to see Solution

The median is the better measure of the “centre” than the mean because 59 of the values are either \$280, 000 or \$315, 000 and only one is \$2, 500, 000. The \$2, 500, 000 is an outlier. Either \$280, 000 or \$315, 000 gives us a better sense of the middle of the data.

Mode

Another measure of the center of the data is the mode. The **mode** is the most frequently occurring value in the set of data. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A set of data can also have no mode if all of the observations in the data are unique.

Unlike the mean and the median, the mode can be calculated for both qualitative data and quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

CALCULATING THE MODE IN EXCEL

To find the mode in Excel:

- Use the **count** and **mode.mult** function to determine the number of modes in the data. Enter **count(mode.mult(array))** into a cell where **array** is the array or cell range containing the data. This function will output the number of modes present in the data.
- If the output from the **count(mode.mult(array))** function is 1, then the data has a single mode. To find the single mode, use the **mode.sngl(array)** function, where **array** is the array or cell range containing the data. The output from the **mode.sngl** function is the value of single mode in the data.
 - Visit the Microsoft page for more information about the **mode.sngl** function.
- If the output from the **count(mode.mult(array))** function is greater than 1, then the data contains multiple modes. To find the multiple modes:
 - Left click on a cell, hold and drag down to highlight a number of vertical cells equal to the number of modes in the data. For example, if there are 4 modes in the data, highlight 4 cells in the vertical array.
 - In the highlighted cells, enter the **mode.mult(array)** function, where **array** is the array or cell range containing the data.
 - After entering the **mode.mult** function in the vertical array, press **CTRL+SHIFT+ENTER**. Because the output from this function is an array, we must press **CTRL+SHIFT+ENTER** (and not **ENTER**) to produce the array output.
 - The output from the **mode.mult** function are the modes in the data.
 - Visit the Microsoft page for more information about the **mode.mult** function.

EXAMPLE

Statistics exam scores for 20 students are as follows:

50	53	59	59	63	63	72	72	72	72
72	76	78	81	83	84	84	84	90	93

Find the mode.

Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A20.

Start by using the **count** function to count the number of modes in the data:

Function	count(mode.mult(...))
Field 1	A1:A20
Answer	1

Because the output from the **count(mode.mult(...))** function is 1, there is only 1 mode in the data. To find the single mode, we use the **mode.sngl** function:

Function	mode.sngl
Field 1	A1:A20
Answer	72

By examining the data, we can see that **72** is the most frequently occurring value (5 times) and that **72** is the only value that occurs 5 times.

TRY IT

The number of books checked out from the library from 25 students are as follows:

0	0	0	1	2
3	3	4	4	5
5	7	7	7	7
8	8	8	9	10
10	11	11	12	12

Find the mode.

Click to see Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A25.

Start by using the **count** function to count the number of modes in the data:

Function	count(mode.mult(...))
Field 1	A1:A25
Answer	1

Because the output from the **count(mode.mult(...))** function is 1, there is only 1 mode in the data. To find the single mode, we use the **mode.sngl** function:

Function	mode.sngl
Field 1	A1:A25
Answer	7

The most frequent number of books is 7, which occurs four times.

EXAMPLE

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3	4	8	8	10	11	12	13	14	15
15	16	16	17	17	18	21	22	22	24
24	25	26	26	27	27	29	29	31	32
33	33	34	34	35	37	40	44	44	47

Calculate the mode.

Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A40.

Start by using the **count** function to count the number of modes in the data:

Function	count(mode.mult(...))
Field 1	A1:A40
Answer	12

Because the output from the **count(mode.mult(...))** function is 12, there are 12 modes in the data. To find the multiple modes, we use the **mode.mult** function. Left-click on a cell, hold and drag down to highlight 12 vertical cells. In the highlighted cells, enter the **mode.mult** function:

Function	mode.mult
Field 1	A1:A40
Answer	8, 15, 16, 17, 22, 24, 26, 27, 29, 33, 34, 44

Because the output from the **mode.mult** function is a (vertical) array after entering the function, press **CTRL+SHIFT+ENTER** (not **ENTER** by itself).

TRY IT

Ten credit scores are

645	680	700	720	517	630	598	739	720	680
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Calculate the mode.

Click to see Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A10.

Start by using the **count** function to count the number of modes in the data:

Function	count(mode.mult(...))
Field 1	A1:A10
Answer	2

Because the output from the **count(mode.mult(...))** function is 2, there are 2 modes in the data. To find the multiple modes, we use the **mode.mult** function. Left click on a cell, hold and drag down to highlight 2 vertical cells. In the highlighted cells, enter the **mode.mult** function:

Function	mode.mult
Field 1	A1:A10
Answer	680, 720

Because the output from the **mode.mult** function is a (vertical) array after entering the function, press **CTRL+SHIFT+ENTER** (not **ENTER** by itself).



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=55#oembed-1>

Video: “Finding mean, median, and mode | Descriptive statistics | Probability and Statistics | Khan Academy” by Khan Academy [3:55] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

When to Use Each Measure of Central Tendency

The measures of central tendency tell us about the center of the data but often give different answers. So, how do we know when to use each? Here are some general rules:

1. The mean is the most frequently used measure of central tendency and is generally considered the best measure of central tendency.
2. Median is the preferred measure of central tendency when:
 - a. There are a few extreme values or outliers in the distribution of the data. (Note: Remember that a single outlier can have a great effect on the mean).
 - b. There are some missing or undetermined values in the data.
 - c. There is an open-ended distribution (For example, if a data field measuring the number of children has options 0, 1, 2, 3, 4, 5 or “6 or more,” then the “6 or more” field is open-ended and makes calculating the mean impossible because we do not know the exact values for this field).
 - d. Data that is measured on an ordinal scale.
3. Mode is the preferred measure when data are measured on a nominal or ordinal scale.

Exercises

1. How much time does it take to travel to work in a particular region? The table below shows the commute time for a sample of workers in the region who are at least 16 years old and do not work at home.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

- Find the mean for this data.
- Find the median for this data.
- Find the mode for this data.

Click to see Answer

- 23.462 minutes
- 23.5 minutes
- 24 minutes, 24.7 minutes, 25.7 minutes, 23.1 minutes, 18.6 minutes, 23.3 minutes, 22.6 minutes, 23.9 minutes, 25.5 minutes, 27 minutes

- The following data shows the lengths, in feet, of a sample of boats moored in a marina.

19	35	29	26	21	40	33	33	34
25	20	37	30	26	23	24	29	16
28	25	20	39	32	27	27	27	17

- Calculate the mean.
- Calculate the median.
- Find the mode.

Click to see Answer

- 27.33 feet
- 27 feet
- 25 feet, 27 feet

- The data below is the weight, in pounds, of all members of a particular NFL team.

177	210	270	275	212	185	200	241	250
220	259	185	210	272	285	212	250	302
205	232	280	285	184	265	215	223	265
260	278	185	228	273	242	185	241	290
210	276	290	206	174	286	247	190	215
245	205	178	290	280	188	230	260	

- Find the mean for this data.
- Find the median for this data.
- Find the mode for this data.

Click to see Answer

- 236.25 pounds
- 241 pounds
- 185 pounds

- A sample of 35 post-secondary institutions was taken from across the U.S. The data below shows the number of students enrolled at each institution.

6,414	1,550	2,109	9,350	21,828	4,300	5,944
5,722	2,825	2,044	5,481	5,200	5,853	10,012
6,357	27,000	9,414	7,681	3,200	17,500	9,200
7,380	18,314	6,557	13,713	17,768	7,493	2,771
2,861	1,263	7,285	28,165	5,080	11,622	2,750

- Calculate the mean for this data.
- Calculate the median for this data.
- Find the mode for this data.
- If you were to build a new community college, which piece of information would be more valuable: the mode or the mean? Explain.

Click to see Answer

- 8,628.74 students
- 6,414 students
- no mode
- The mean because there is no mode in this data.

5. Forty randomly selected students were asked the number of pairs of sneakers they owned. The data is recorded below

1	1	2	2	2	2	2	3
3	3	3	3	3	3	3	4
4	4	4	4	4	4	4	4
4	4	4	5	5	5	5	5
5	5	5	5	5	5	5	7

- Calculate the mean for this data.
- Calculate the median for this data.
- Find the mode for this data.

Click to see Answer

- 3.775 sneakers
- 4 sneakers
- 4 sneakers, 5 sneakers

6. The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.
- What does it mean for the median age to rise?
 - Give two reasons why the median age could rise.

Click to see Answer

- In 1980, half of the population was younger than 30 years of age, and the other half of the population was older than 30 years of age. In 1991, half of the population was older than 33.1 years of age, and the other half of the population was older than 33.1 years of age. Because the median age rose over that decade, it means that the overall age of the population rose over that time period.
- The median age could rise because fewer children are being born and because people are living longer.

Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

2.3 SKEWNESS AND THE MEAN, MEDIAN, AND MODE

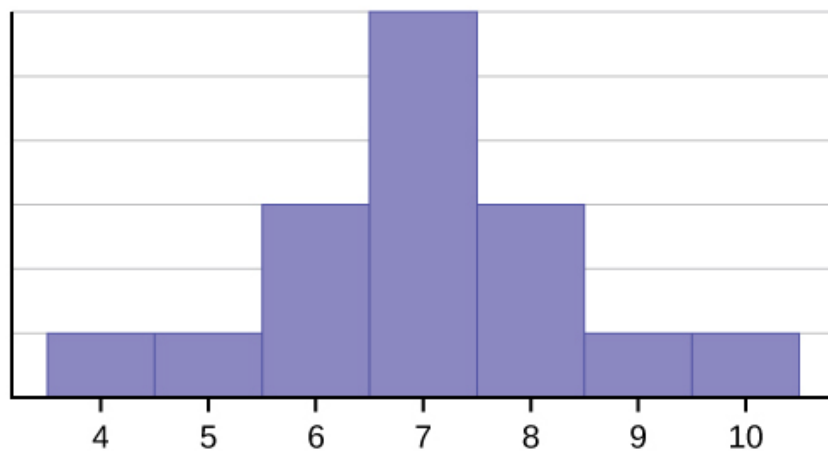
LEARNING OBJECTIVES

- Identify the shape of a set of data.

Consider the following data set:

4	5	6	6	6	7	7	7
7	7	7	8	8	8	9	10

This data set can be represented by the following histogram. Each interval has a width of one, and each value is located in the middle of an interval.



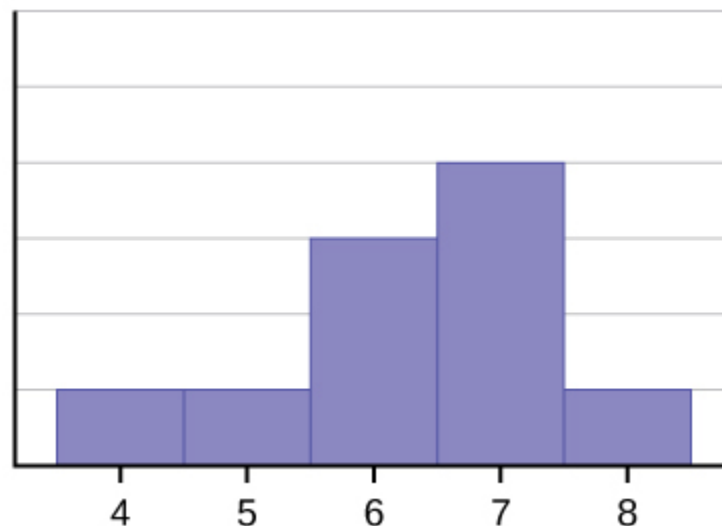
The histogram above displays a **symmetrical** distribution of data. A distribution is symmetrical if a

vertical line can be drawn at some point in the histogram so that the shape to the left and to the right of the vertical line are mirror images of each other. For the above data set, the mean, the median, and the mode are each seven. In a perfectly symmetrical distribution, the mean and the median are the same. This example has one mode, and the mode is the same as the mean and median. In a symmetrical distribution that has multiple modes, the modes would be different from the mean and median.

Consider the following data set:

4	5	6	6	6	7	7	7	7	8
---	---	---	---	---	---	---	---	---	---

This data set can be represented by the following histogram. Each interval has a width of one, and each value is located in the middle of an interval.

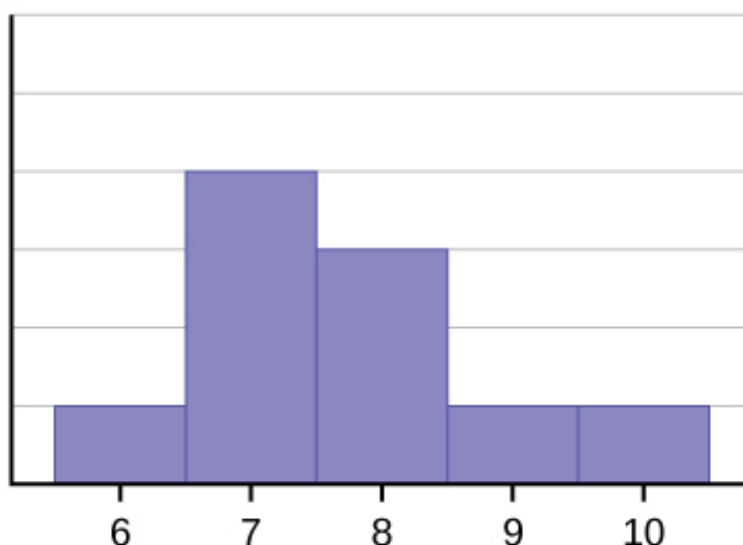


The histogram above is not symmetrical. The left-hand side seems “chopped off” compared to the right side. A distribution of this type is called **skewed to the left** because it is pulled out to the left. The mean of this data is 6.3, the median is 6.5, and the mode is 7. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

Consider the following data set:

6	7	7	7	7	8	8	8	9	10
---	---	---	---	---	---	---	---	---	----

This data set can be represented by the following histogram. Each interval has a width of one, and each value is located in the middle of an interval.



The histogram above is also not symmetrical. In this case, the data is **skewed to the right**. The mean for this data is 7.7, the median is 7.5, and the mode is 7. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize:

- If the distribution of the data is symmetrical, then $\text{mean} = \text{median} = \text{mode}$ (assuming there is only one mode). If there are multiple modes in a symmetric distribution, the modes would be different from the mean and the median, but the mean and median would still be equal.
- If the distribution of the data is skewed to the left, then $\text{mean} < \text{median} < \text{mode}$.
- If the distribution of the data is skewed to the right, then $\text{mean} > \text{median} > \text{mode}$.

Skewness and symmetry become important when we discuss probability distributions in later chapters.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=63#oembed-1>

Video: “Elementary Business Statistics | Skewness and the Mean, Median, and Mode” by Janux

[3:58] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

EXAMPLE

Statistics are used to compare and sometimes identify authors. The following list shows a simple random sample that compares the letter counts for three authors.

Terry									
7	9	3	3	3	4	1	3	2	2
Davis									
3	3	3	4	1	4	3	2	3	1
Maris									
2	3	4	4	4	6	6	6	8	3

1. Make a dot plot for the three authors and compare the shapes.
2. Calculate the mean for each.
3. Calculate the median for each.
4. Describe any pattern between the shape and the measures of centre.

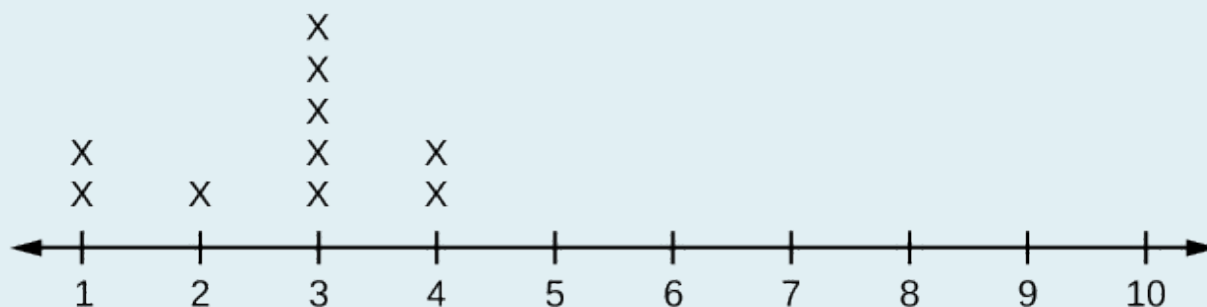
Solution

Terry's Letter Count



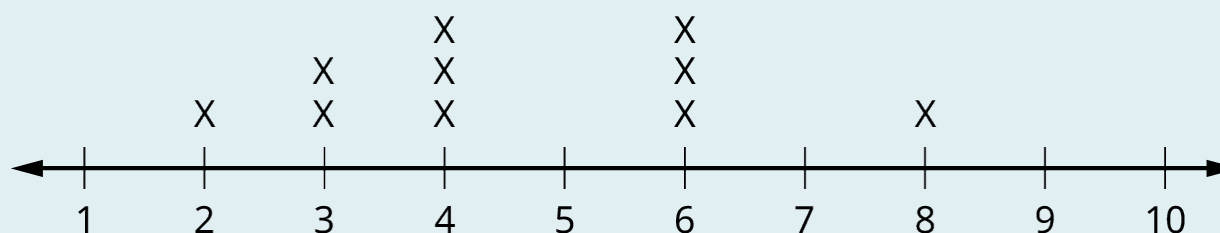
Terry's distribution has a right (positive) skew.

Davi's Letter Count



Davis' distribution has a left (negative) skew

Mari's Letter Count



Maris' distribution is symmetrically shaped.

1. Terry's mean is **3.7**, Davis' mean is **2.7**, Maris' mean is **4.6**.
2. Terry's median is **3**, Davis' median is **3**, Maris' median is **4**.
3. It appears that the median is always closest to the high point (the mode), while the mean tends to be farther out on the tail. In a symmetrical distribution, the mean and the median are both

centrally located close to the high point of the distribution.

Exercises

1. State whether the data are symmetrical, skewed to the left, or skewed to the right.

a.

16	17	19	22	22	22	22	22	23
----	----	----	----	----	----	----	----	----

b.

87	87	87	87	87	88	89	89	90	91
----	----	----	----	----	----	----	----	----	----

Click to see Answer

- a. Left-skewed because the mean (20.56) is less than the median (22).
- b. Right-skewed because the mean (88.2) is greater than the median (87.5) is greater than the mode (87).

2. When the data are skewed left, what is the typical relationship between the mean and median?

Click to see Answer

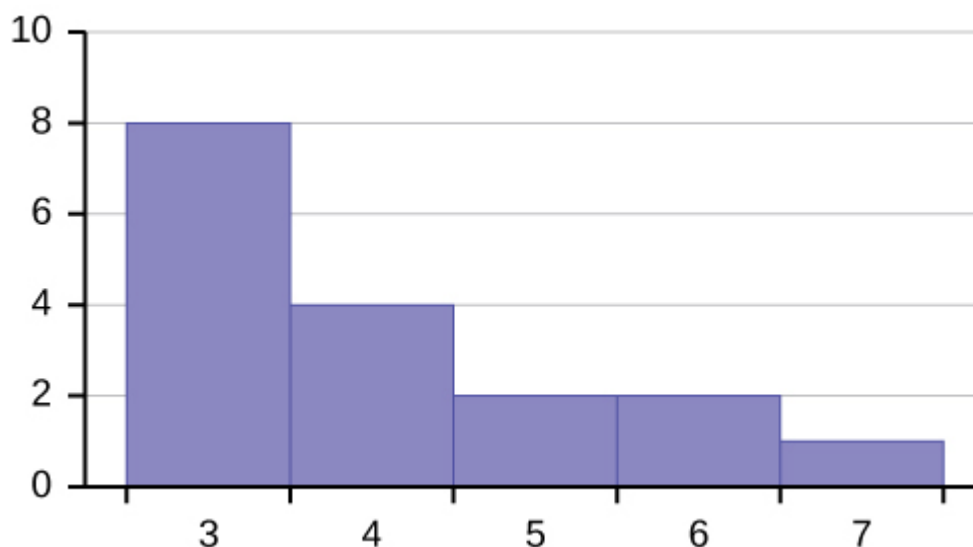
mean < median < mode

3. When the data are symmetrical, what is the typical relationship between the mean and median?

Click to see Answer

mean = median = mode

4. Consider the following distribution.

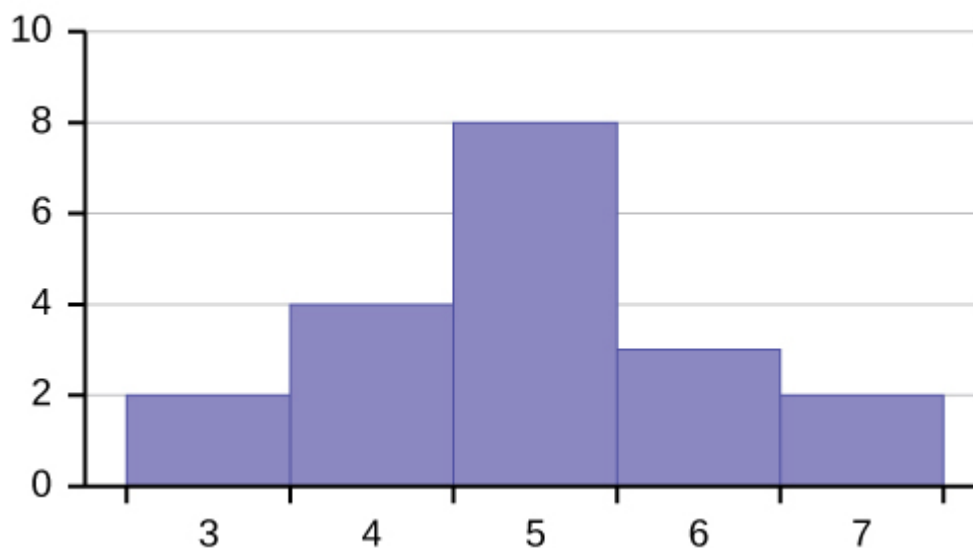


- Describe the shape of this distribution.
- Describe the relationship between the mode and the median of this distribution.
- Describe the relationship between the mean and the median of this distribution.

Click to see Answer

- Right-skewed.
- The mode is less than the median.
- The median is less than the mean.

5. Consider the following distribution.



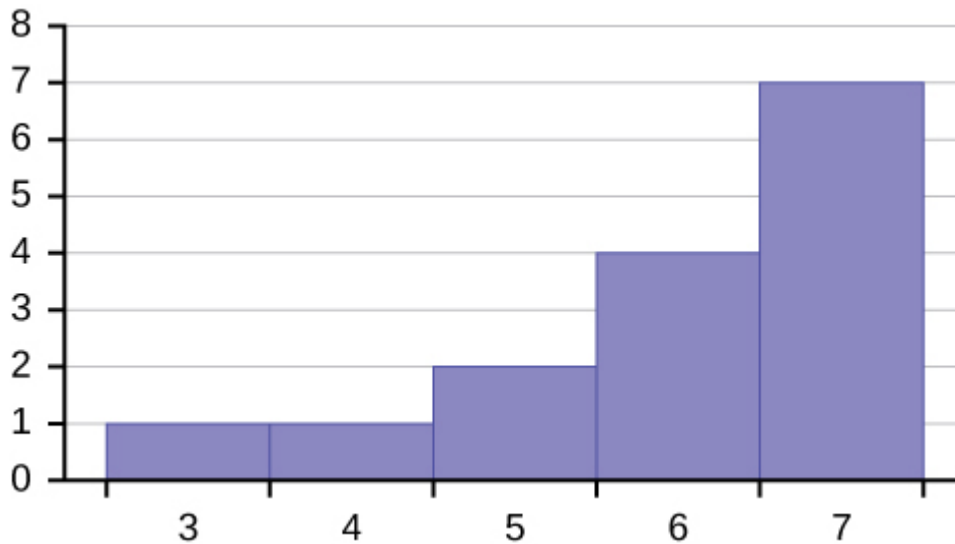
- Describe the relationship between the mode and the median of this distribution.
- Are the mean and the median equal in this distribution? Why or why not?

Click to see Answer

- The mode and the median are equal.

- b. No, because the distribution is not symmetric.

6. Consider the following distribution.



- Describe the shape of this distribution.
- Describe the relationship between the mode and the median of this distribution.
- Describe the relationship between the mean and the median of this distribution.

Click to see Answer

- Left-skewed.
- The median is less than the mode.
- The mean is less than the median.

7. The mean and median for the data shown below are the same. Is the data perfectly symmetrical? Why or why not?

3	4	5	5	6
6	6	6	7	7
7	7	7	7	7

Click to see Answer

The data is not perfectly symmetrical. The mean and the median are both 6, but the mode is 7.

8. Of the three measures, which tends to reflect skewing the most: the mean, the mode, or the

median? Why?

Click to see Answer

The mean because the mean is the measure of central tendency that is most susceptible to extreme values.

9. In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

Click to see Answer

When there are multiple modes in the data.

“2.4 Skewness and the Mean, Median, and Mode” and “2.7 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

2.4 MEASURES OF POSITION

LEARNING OBJECTIVES

- Recognize, describe, calculate, and interpret the measures of the position of data: quartiles and percentiles.

The common measures of position are **quartiles** and **percentiles**. Previously, we learned that the **median** is a number that measures the “center” of the data. But the median can also be thought of as a measure of position because the median is the “middle value” of a set of data. The median is a number that separates ordered data into halves. Half of the values in the data are the same number or smaller than the median, and half of the values in the data are the same number or larger.

For example, consider the following data, already ordered from smallest to largest:

1	1	2	2	4	6	6.8
7.2	8	8.3	9	10	10	11.5

Because there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two:

$$\frac{6.8 + 7.2}{2} = 7$$

The median of this data is 7. We can see that half (or 50\%) of the values are less than seven and half (or 50\%) of the values are larger than seven.

The median is an example of both a quartile and a percentile. The median is also the second quartile, Q_2 , and the 50th percentile, P_{50} .

Quartiles

Quartiles are numbers that separate the data into quarters (four parts). Like the median, quartiles may or may not be an actual value in the set of data. To find the quartiles, order the data (from smallest to largest) and then find the median or second quartile. The first quartile, Q_1 , is the middle value of the lower half of the data, and the third quartile, Q_3 , is the middle value of the upper half of the data. To get the idea, consider the same (ordered) data set used above:

1	1	2	2	4	6	6.8
7.2	8	8.3	9	10	10	11.5

The median or **second quartile** is 7. The lower half of the data are:

1	1	2	2	4	6	6.8
---	---	---	---	---	---	-----

The middle value of the lower half of the data is 2. The number 2, which is part of the data, is the **first quartile**, Q_1 . One-fourth (or 25\%) of the values in the data are the same as or less than 2, and three-fourths (or 75\%) of the values are more than 2.

The upper half of the data are:

7.2	8	8.3	9	10	10	11.5
-----	---	-----	---	----	----	------

The middle value of the upper half of the data is 9. The **third quartile**, Q_3 , is 9. Three-fourths (or 75\%) of the values in the data are the same or less than 9. One-fourth (or 25\%) of the values in the data set are greater than 9.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50\% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$$IQR = Q_3 - Q_1$$

The IQR can help to determine potential **outliers**. A value is suspected to be a potential outlier if it is less than $1.5 \times IQR$ below the first quartile or more than $1.5 \times IQR$ above the third quartile. Potential outliers always require further investigation.

NOTE

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors, some kind of abnormality, or they may be a key to understanding the data.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=65#oembed-1>

Video: “Median, Quartiles and Interquartile Range : ExamSolutions” by ExamSolutions [12:36] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

CALCULATING QUARTILES IN EXCEL

To find quartiles in Excel, use the **quartile.exc(array, quartile number)** function.

- For **array**, enter the array or cell range containing the data.
- For **quartile number**, enter the quartile (1, 2 or 3) being calculated.

The output from the **quartile.exc** function is the value of the corresponding quartile. For example, **quartile.exc(array,1)** returns the value of the first quartile where 25\% of the observations in the data are less than or equal to the value of the first quartile.

Visit the Microsoft page for more information about the **quartile.exc** function.

NOTE

We are using the **quartile.exc** function, and not the **quartile.inc** function. These two functions calculate the quartiles in slightly different ways.

- The **quartile.exc** function calculates the quartiles by first finding the median of the data set. The first quartile is the median of the lower half of the data, excluding the median value from the lower half of the data. The third quartile is the median value of the upper half of the data, excluding the median value from the upper half of the data.
- The **quartile.inc** function calculates the quartiles by first finding the median of the data set. The first quartile is the median of the lower half of the data, including the median value (if the median is a number in the data set) with the lower half of the data. The third quartile is the median of the upper half of the data, excluding the median value (if the median is a number in the data set) with the upper half of the data.

In some cases, the **quartile.exc** and **quartile.inc** will return the same values, depending on whether or not the median is a number in the data set.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=65#oembed-2>

Video: “How To Find Quartiles and Construct a Boxplot in Excel” by Joshua Emmanuel [4:13] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

EXAMPLE

For the following 13 real estate prices, calculate the three quartiles and the *IQR*. Determine if any prices are potential outliers. The prices are in dollars.

389,950	230,500	158,000	479,000	639,000	114,950	5,500,000
387,000	659,000	529,000	575,000	488,800	1,095,000	

Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A13.

For the first quartile Q_1 :

Function	quartile.exc
Field 1	A1:A13
Field 2	1
Answer	\$308,750

For the second quartile Q_2 :

Function	quartile.exc
Field 1	A1:A13
Field 2	2
Answer	\$488,800

For the third quartile Q_3 :

Function	quartile.exc
Field 1	A1:A13
Field 2	3
Answer	\$649,000

For the *IQR*: $IQR = 649,000 - 308,750 = \$340,250$

To determine if there are any outliers:

$$1.5 \times IQR = 1.5 \times 340,250 = 510,375$$

$$Q_1 - 1.5 \times IQR = 308,750 - 510,375 = -201,625$$

$$Q_3 + 1.5 \times IQR = 649,000 + 510,375 = 1,159,375$$

No house price is less than $-\$201,625$. However, $\$5,500,000$ is more than $\$1,159,375$. Therefore, $\$5,500,000$ is a potential **outlier**.

NOTE

Quartiles have the same units as the data. In this case, the data is measured in dollars, so the quartiles are also in dollars.

TRY IT

For the following 11 salaries, calculate the three quartiles and the *IQR*. Are any of the salaries outliers? The salaries are in dollars.

33,000	72,000	54,000
64,500	68,500	120,000
28,000	69,000	40,500
54,000	42,000	

Click to see Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A11.

For the first quartile Q_1 :

Function	quartile.exc
Field 1	A1:A11
Field 2	1
Answer	\$40,500

For the second quartile Q_2 :

Function	quartile.exc
Field 1	A1:A11
Field 2	2
Answer	\$54,000

For the third quartile Q_3 :

Function	quartile.exc
Field 1	A1:A11
Field 2	3
Answer	\$69,000

For the *IQR*: $IQR = 69,000 - 40,500 = \$28,500$

To determine if there are any outliers:

$$1.5 \times IQR = 1.5 \times 28,500 = 42,750$$

$$Q_1 - 1.5 \times IQR = 40,500 - 42,750 = -2,250$$

$$Q_3 + 1.5 \times IQR = 69,000 + 42,750 = 111,750$$

No salary is less than $-\$2,250$. However, $\$120,000$ is more than $\$111,750$, so $\$120,000$ is a potential outlier.

TRY IT

Find the interquartile range for the following two data sets and compare them.

Test Scores for Class A									
69	96	81	79	65	76	83	99	89	67
90	77	85	98	66	91	77	69	80	94

Test Scores for Class B									
90	72	80	92	90	97	92	75	79	68
70	80	99	95	78	73	71	68	95	100

Click to see Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data for Class A into column A from cell A1 to A20 and the data for Class B into column B from cell B1 to B20.

Class A

For the first quartile Q_1 :

Function	quartile.exc
Field 1	A1:A20
Field 2	1
Answer	70.75

For the third quartile Q_3 :

Function	quartile.exc
Field 1	A1:A20
Field 2	3
Answer	90.75

For the IQR : $IQR = 90.75 - 70.75 = 20$

Class B

For the first quartile Q_1 :

Function	quartile.exc
Field 1	B1:B20
Field 2	1
Answer	72.25

For the third quartile Q_3 :

Function	quartile.exc
Field 1	B1:B20
Field 2	3
Answer	94.25

For the IQR : $IQR = 94.25 - 72.25 = 22$

The data for Class B has a larger IQR , so the scores between Q_3 and Q_1 (the middle 50\% of the data) for the data for Class B are more spread out and not clustered about the median.

Percentiles

Percentiles are numbers that separate the (ordered) data into hundredths (100 parts). Like quartiles, percentiles may or may not be part of the data. The n th percentile, P_n , is the value where $n\%$ of the observations in the data are less than or equal to the value of the n th percentile. To score in the 90th percentile of an exam does not mean, necessarily, that the student received 90\% on a test. The 90th percentile means that 90\% of test scores are less than or equal to the student's score and 10\% of the test scores are the same or greater than the student's score. Percentiles are mostly used with very large data sets.

Quartiles are special percentiles. The first quartile, Q_1 , is the same as the 25th percentile, and the third quartile, Q_3 , is the same as the 75th percentile. The median is the 50th percentile.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into an SAT score of at least 1220.

CALCULATING PERCENTILES IN EXCEL

To find the k th percentiles in Excel, use the **percentile.exc(array, percent)** function.

- For **array**, enter the array or cell range containing the data.
- For **percent**, enter the percentile (as a decimal) being calculated. For example, if we are calculating the 60th percentile, we would enter 0.6 for the percent in the **percentile.exc** function.

The output from the **percentile.exc** function is the value of the corresponding percentile. For example, **percentile.exc(array,0.6)** returns the value of the 60th percentile where 60% of the observations in the data are less than or equal to the value of the 60th percentile.

Visit the Microsoft page for more information about the **percentile.exc** function.

NOTE

We are using the **percentile.exc** function, and not the **percentile.inc** function. Like the quartile functions, the **percentile.exc** and **percentile.inc** calculate the percentiles in different ways, and so give slightly different answers for the percentiles.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=65#oembed-3>

Video: “Percentiles – How to calculate Percentiles, Quartiles, ...” by Joshua Emmanuel [3:44] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

EXAMPLE

Listed are twenty-nine ages (in years) for trees found in the Saint Louis Botanical Garden.

18	21	22	25	26	27	29	30	31	33
36	37	41	42	47	52	55	57	58	62
64	67	69	71	72	73	74	76	77	

1. Find the 70th percentile.
2. Find the 83rd percentile.

Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A29.

For the 70th percentile P_{70} :

Function	percentile.exc
Field 1	A1:A29
Field 2	0.7
Answer	64 years

For the 83rd percentile P_{83} :

Function	percentile.exc
Field 1	A1:A29
Field 2	0.83
Answer	71.9 years

NOTE

Percentiles have the same units as the data. In this case, the data is measured in years, so the percentiles are also in years.

TRY IT

Listed are **29** ages (in years) for Academy Award-winning best actors.

18	21	22	25	26	27	29	30	31	33
36	37	41	42	47	52	55	57	58	62
64	67	69	71	72	73	74	76	77	

Calculate the **20th** percentile and the **55th** percentile.

Click to see Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A29.

For the 20th percentile P_{20} :

Function	percentile.exc
Field 1	A1:A29
Field 2	0.2
Answer	27 years

For the 55th percentile P_{55} :

Function	percentile.exc
Field 1	A1:A29
Field 2	0.55
Answer	53.5 years

Interpreting Percentiles and Quartiles

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the value of the n th percentile. For example, 15\% of the data values are less than or equal to the value of the 15th percentile. Note that low percentiles always correspond to lower data values, and high percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is “good” or “bad.” The interpretation of whether a certain percentile is “good” or “bad” depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered “good,” but in other contexts, a high percentile might be considered “good”. In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles or quartiles properly is important not only when describing data but also when calculating probabilities in later chapters of this text. When writing the interpretation of a percentile or quartile in the context of the given data, the sentence should contain the following information:

- Information about the context of the situation being considered,
- The data value (value of the variable) that represents the percentile/quartile.

- The percent of individuals or items with data values less than or equal to the percentile/quartile.

EXAMPLE

On a timed math test, the first quartile for the time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Solution

- Interpretation: 25\% of students finished the exam in less than or equal to 35 minutes.
- In this context, a low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If a student takes too long, they might not be able to finish.)

TRY IT

For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Click to see Solution

- Interpretation: 75\% of runners finished the race in less than or equal to 11.5 seconds.
- In this context, a lower percentile is good because finishing a race more quickly is desirable.

EXAMPLE

On a 20 question math test, the 70th percentile for the number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

Solution

- Interpretation: 70\% of students answered less than or equal to 16 questions correctly.

TRY IT

On a 60 point written assignment, the 80th percentile for the number of points earned was 49. Interpret the 80th percentile in the context of this situation.

Click to see Solution

- Interpretation: 80\% of students earned less than or equal to 49 points.

EXAMPLE

At a community college, it was found that the 30th percentile of credit units that students are enrolled for is 7 units. Interpret the 30th percentile in the context of this situation.

Solution

- Interpretation: 30\% of students are enrolled in less than or equal to 7 credit units.
- In this context, there is no “good” or “bad” value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

TRY IT

During a season, the 40th percentile for points scored per player in a game is 8. Interpret the 40th percentile in the context of this situation.

Click to see Solution

- Interpretation: 40\% of players scored less than or equal to 8 points.

Exercises

1. How much time does it take to travel to work in a particular region? The table below shows

the commute time for a sample of workers in the region who are at least 16 years old and do not work at home.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

- Find the first quartile.
- Interpret the first quartile.
- Find the third quartile.
- The middle 50\% of the travel times lie between what two numbers?
- Interpret the third quartile.
- Find the *IQR*.
- Find the 45th percentile.
- Interpret the 45th percentile.

Click to see Answer

- 21.475 minutes
- 25\% of the workers take at most 21.475 minutes to travel to work.
- 25.55 minutes
- 21.475 minutes to 25.55 minutes
- 75\% of the workers take at most 25.55 minutes to travel to work.
- 4.075 minutes
- 23.29 minutes
- 45\% of the workers take at most 23.29 minutes to travel to work.

- The following data shows the lengths, in feet, of a sample of boats moored in a marina.

19	35	29	26	21	40	33	33	34
25	20	37	30	26	23	24	29	16
28	25	20	39	32	27	27	27	17

- Find the first quartile.
- Interpret the first quartile.

- c. Find the third quartile.
- d. Interpret the third quartile.
- e. The middle 50\% of boat lengths lie between what two numbers?
- f. Find the *IQR*.
- g. Find the 63rd percentile.
- h. Interpret the 63rd percentile.

Click to see Answer

- a. 23 feet
- b. 25\% of the boats are less than or equal to 23 feet in length.
- c. 33 feet
- d. 23 feet to 33 feet
- e. 75\% of the boats are less than or equal to 33 feet in length.
- f. 10 feet
- g. 28.64 feet
- h. 63\% of the boats are less than or equal to 28.64 feet in length.

3. The data below is the weight, in pounds, of all members of a particular NFL team.

177	210	270	275	212	185	200	241	250
220	259	185	210	272	285	212	250	302
205	232	280	285	184	265	215	223	265
260	278	185	228	273	242	185	241	290
210	276	290	206	174	286	247	190	215
245	205	178	290	280	188	230	260	

- a. Find the first quartile.
- b. Interpret the first quartile.
- c. Find the third quartile.
- d. Interpret the third quartile.
- e. The middle 50\% of the player's weights lie between what two numbers?
- f. Find the *IQR*.
- g. Find the 87th percentile.
- h. Interpret the 87th percentile.

Click to see Answer

- a. 205.5 pounds

- b. 25\% of the football players weigh less than or equal to 205.5 pounds.
- c. 272.5 pounds
- d. 75\% of the football players weigh less than or equal to 272.5 pounds.
- e. 205.5 pounds to 272.5 pounds
- f. 67 pounds
- g. 284.9 pounds
- h. 87\% of football players weigh less than or equal to 284.9 pounds.

4. A sample of 35 post-secondary institutions was taken from across the U.S. The data below shows the number of students enrolled at each institution.

6,414	1,550	2,109	9,350	21,828	4,300	5,944
5,722	2,825	2,044	5,481	5,200	5,853	10,012
6,357	27,000	9,414	7,681	3,200	17,500	9,200
7,380	18,314	6,557	13,713	17,768	7,493	2,771
2,861	1,263	7,285	28,165	5,080	11,622	2,750

- a. Find the first quartile.
- b. Interpret the first quartile.
- c. Find the third quartile.
- d. Interpret the third quartile.
- e. Find the *IQR*.
- f. Find the 65th percentile.
- g. Interpret the 65th percentile.

Click to see Answer

- a. 3,200 students
- b. 25\% of the colleges enroll less than or equal to 3,200 students.
- c. 10,012 students
- d. 75\% of the colleges enroll less than or equal to 10,012 students.
- e. 6,812 students
- f. 8,288.6 students
- g. 65\% of the colleges enroll less than or equal to 8,288.6 students.

5. Forty randomly selected students were asked the number of pairs of sneakers they owned.

The data is recorded below

1	1	2	2	2	2	2	3
3	3	3	3	3	3	3	4
4	4	4	4	4	4	4	4
4	4	4	5	5	5	5	5
5	5	5	5	5	5	5	7

- Find the first quartile.
- Interpret the first quartile.
- Find the third quartile.
- Interpret the third quartile.
- Find the *IQR*.
- Find the 90th percentile.
- Interpret the 90th percentile.

Click to see Answer

- 3 pairs of sneakers
 - 25\% of the students own at most 3 pairs of sneakers.
 - 5 pairs of sneakers
 - 75\% of the students own at most 5 pairs of sneakers.
 - 2 pairs of sneakers
 - 5 pairs of sneakers
 - 90\% of the students own at most 5 pairs of sneakers.
- 6.
- For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
 - The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
 - A bicyclist in the 90th percentile of a bicycle race completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.
 - For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
 - The 40th percentile of speeds in a particular race is 7.5 kilometres per hour. Write a

sentence interpreting the 40th percentile in the context of the situation.

Click to see Answer

- a. A low percentile is more desirable because it means that the runner completed the race with a low time and ran a faster race.
 - b. 20\% of the runners had run times of 5.2 minutes or less.
 - c. The cyclist is among the slowest. In this context, beginning in a high percentile means the cyclist had a high race completion time. 90\% of the racers completed the race in 1 hour and 12 minutes or less.
 - d. A high percentile is more desirable because it means that the runner had a high speed and ran a faster race.
 - e. 40\% of the speeds in the race are less than or equal to 7.5 kilometres per hour.
7. On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Click to see Answer

A high percentile because it means a higher grade.

8. Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

Click to see Answer

Because Mina's wait time is in a high percentile, it means that Mina's wait time is among the longest. 85\% of the wait times are 32 minutes or less.

9. In a study collecting data about the repair costs of damage to automobiles in certain types of crash tests, a certain model of car had \$1,700 in damage and was in the 90th percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90th percentile in the context of this problem.

Click to see Answer

Because the cost is in a high percentile, it means that this model of car is among the most expensive to repair. 90\% of the cars have damage costs of \$1,700 or less.

10. Suppose that you want to buy a house. You and your realtor have determined that the most expensive house you can afford is in the 34th percentile. The 34th percentile of housing prices in the town you want to move to is \$240,000. In this town, can you afford 34% of the houses or 66% of the houses? Explain.

Click to see Answer

34% because 34% of the houses cost \$240,000 or less.

11. Using the number of full-time equivalent students (FTES) each year at a local college for the past 40 years, the first quartile is 528.5 FTES and the third quartile is 1,447.5 FTES.
- 75% of all years have an FTES at or below what value?
 - 75% of all years have an FTES above what value?
 - What percent of the FTES were from 528.5 to 1447.5? How do you know?
 - What is the *IQR*? What does the *IQR* represent?

Click to see Answer

- 1,447.5 FTES
- 528.5 FTES
- 50%
- 919 FTES. This is the spread of the middle 50% of the data.

“2.5 Measures of Location” and “2.7 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

2.5 MEASURES OF VARIABILITY

LEARNING OBJECTIVES

- Recognize, describe, calculate, and analyze the measures of the spread of data: variance, standard deviation, and range.

It can be misleading to only use the measures of central tendency (mean, median, mode) to describe a data set. Measures of central tendency describe the center of a distribution. Measures of dispersion or variability are used to describe the spread or dispersion of the data. So far in this chapter, we have already seen a measure of variability—the interquartile range. The interquartile range describes the spread of the middle 50\% of the data. But there are other measures of variability, including range, variance, and standard deviation.

Range

The **range** is the difference between the largest and smallest value in a set of data:

$$\text{Range} = \text{Maximum Value} - \text{Minimum Value}$$

Range is a poor measure of variability because it is based on only two values in the data set (the largest and smallest values) and is highly influenced by outliers. Also, the range does not help us distinguish between two data sets with the same largest and smallest values because the two data sets will have the same range.

EXAMPLE

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows:

3	4	8	8	10	11	12	13	14	15
15	16	16	17	17	18	21	22	22	24
24	25	26	26	27	27	29	29	31	32
33	33	34	34	35	37	40	44	44	47

Calculate the range.

Solution

The largest value is 47 and the smallest value is 3, so

$$\text{Range} = 47 - 3 = 44 \text{ months}$$

Variance and Standard Deviation

An important characteristic of any set of data is the variation in the data from the mean. In some data sets, the data values are concentrated close to the mean, but in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures, on average, how far data values are from their mean. The standard deviation provides a numerical measure of the overall amount of variation in a data set and can be used to determine whether a particular data value is close to or far away from the mean.

The standard deviation provides a measure of the overall variation in a data set. The standard deviation is always a non-negative number. The standard deviation is small when the data are all concentrated close to the mean because there is little variation or spread in the data. The standard deviation is larger when the data values are more spread out from the mean because there is a lot

of variation in the data. The lowercase letter s represents the sample standard deviation, and the Greek letter σ represents the population standard deviation.

Suppose that we are studying the amount of time customers wait in line at the checkout at Supermarket A and Supermarket B. The mean wait time at both supermarkets is five minutes. At supermarket A, the standard deviation for the wait time is two minutes and at supermarket B, the standard deviation for the wait time is four minutes. Because supermarket B has a higher standard deviation, we know that there is more variation in the wait times at supermarket B. Overall, wait times at supermarket B are more spread out from the mean, and wait times at supermarket A are more concentrated near the mean.

As well, the standard deviation can be used to determine whether a data value is close to or far from the mean. For example, suppose that Rosa and Binh both shop at Supermarket A, where the mean wait time at the checkout is five minutes, and the standard deviation is two minutes. Suppose Rosa's wait time is seven minutes and Binh's wait time is one minute.

- Rosa's wait time of seven minutes is **two minutes longer than the mean** of five minutes. Because two minutes is equal to one standard deviation, Rosa's wait time of seven minutes is **one standard deviation above the mean** of five minutes.
- Binh's wait time of one minute is **four minutes less than the mean** of five minutes. Because four minutes is equal to two standard deviations, Binh's wait time of one minute is **two standard deviations below the mean** of five minutes.

A data value that is two standard deviations from the mean is just on the borderline for what many statisticians would consider to be far from the mean. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate “rule of thumb” than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations.

Calculating the Standard Deviation

If x is a number, then the difference “ $x - \text{mean}$ ” is called its **deviation from the mean**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols, a deviation is $x - \mu$. For sample data, in symbols, a deviation is $x - \bar{x}$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar but **not** identical. Therefore,

the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lowercase letter s represents the sample standard deviation, and the Greek letter σ represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ .

To calculate the standard deviation, we need to calculate the variance first. The **variance** is the **average of the squares of the deviations** (the $x - \bar{x}$ values for a sample or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance, and the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance, and the sample standard deviation s is the square root of the sample variance. The standard deviation can be thought of as a special average of the deviations.

The formula for the population standard deviation is: $\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$. To calculate a population standard deviation σ :

1. Calculate the deviation from the mean for each data value x : $x - \mu$.
2. Square each of the deviations: $(x - \mu)^2$.
3. Add up the squares of the deviations from the mean calculated in step 2.
4. Divide the sum in step 3 by the population size N .
5. The population standard deviation is the square root of the value from step 4.

The formula for the population variance is $\sigma^2 = \frac{\sum(x - \mu)^2}{N}$. The population variance is the value found in step 4 in the above population standard deviation calculation.

The formula for the sample standard deviation is: $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$. To calculate a sample standard deviation s :

1. Calculate the deviation from the mean for each data value x : $x - \bar{x}$.
2. Square each of the deviations: $(x - \bar{x})^2$.
3. Add up the squares of the deviations from the mean calculated in step 2.
4. Divide the sum in step 3 by the sample size minus 1: $n - 1$.
5. The population standard deviation is the square root of the value from step 4.

The formula for the sample variance is $s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$. The sample variance is the value found in step 4 in the above sample standard deviation calculation.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=67#oembed-2>

Video: “How to calculate Standard Deviation and Variance” by statisticsfun [5:05] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

CALCULATING VARIANCE IN EXCEL

To find the variance in Excel:

- If the data is population data, use the **var.p(array)** function, where **array** is the array or cell range containing the data. The output from the **var.p** function is the population variance.
 - Visit the Microsoft page for more information about the **var.p** function.
- If the data is sample data, use the **var.s(array)** function where **array** is the array or cell range containing the data. The output from the **var.s** function is the sample variance.
 - Visit the Microsoft page for more information about the **var.s** function.

NOTE

There are two different functions to calculate variance in Excel because variance is calculated differently depending on whether the data is from a sample or from a population. When calculating variance, make sure to use the correct function based on the type of data (sample or population).

CALCULATING STANDARD DEVIATION IN EXCEL

To find the standard deviation in Excel:

- If the data is population data, use the **stdev.p(array)** function where **array** is the array or cell range containing the data. The output from the **stdev.p** function is the population standard deviation.
 - Visit the Microsoft page for more information about the **stdev.p** function.
- If the data is sample data, use the **stdev.s(array)** function where **array** is the array or cell range containing the data. The output from the **stdev.s** function is the sample standard deviation.
 - Visit the Microsoft page for more information about the **stdev.s** function.

NOTE

There are two different functions to calculate standard deviation in Excel because standard deviation is calculated differently depending on whether the data is from a sample or from a population. When calculating standard deviation, make sure to use the correct function based on the type of data (sample or population).



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=67#oembed-1>

Video: “Range, Variance, Standard Deviation in Excel” by Joshua Emmanuel [1:11] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

EXAMPLE

In a fifth-grade class, the teacher was interested in the standard deviation of the ages of her students. The following data are the ages, in years, for a sample of 20 fifth-grade students. The ages are rounded to the nearest half year:

9	9.5	9.5	10	10	10	10	10.5	10.5	10.5
10.5	11	11	11	11	11	11	11.5	11.5	11.5

Calculate the mean, the variance, and the standard deviation of the ages of the students. Interpret the standard deviation.

Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A20.

For the mean:

Function	average
Field 1	A1:A20
Answer	10.525 years

For the variance:

Function	var.s
Field 1	A1:A20
Answer	0.5125 years ²

For the standard deviation:

Function	stdev.s
Field 1	A1:A20
Answer	0.7159 years

Interpreting the standard deviation:

On average, the age of any fifth grader is **0.7159** years away from the mean of **10.525** years.

NOTES

1. We are using the **var.s** (not **var.p**) and **stdev.s** (not **stdev.p**) functions to calculate the variance and the standard deviation because the data is from a sample.
2. Standard deviation has the same units as the data. In this case, the data is measured in years, so the standard deviation is also in years.
3. Because the values being added up in the variance calculation are squared, the units of variance are squared units. In particular, the units of variance are the squared units of the data. In this example, the data is measured in years, so the units of the variance are (years)². Because the units of variance are squared units, it can be difficult to intuitively interpret the meaning of the variance.

TRY IT

On a baseball team, the ages, in years, of each of the players are as follows:

21	21	22	23	24
24	25	25	28	29
28	31	32	33	33
34	35	36	36	36
36	38	38	38	40

Find the mean and standard deviation.

Click to see Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A25.

For the mean:

Function	average
Field 1	A1:A25
Answer	30.64 years

For the standard deviation:

Function	stdev.p
Field 1	A1:A25
Answer	5.99 years

NOTE

We are using the **stdev.p** (not **stdev.s**) function to calculate the standard deviation here because the baseball team is a population.

NOTE

Concentrate on what the standard deviation tells us about the data. The standard deviation is a number that measures how far the data is spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ , is a non-negative number. When the standard deviation is zero, there is no dispersion about the mean—that is, all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean and is larger when the data values show more variation from the mean. When the standard deviation is significantly larger than zero, the data values are very spread out about the mean. Outliers in the data can make the standard deviation very large.

The standard deviation, when first presented, can seem unclear. By graphing the data, we can get a better “feel” for the deviations and the standard deviation. In symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph the data.**

EXAMPLE

Use the following sample of exam scores from Susan Dean’s spring pre-calculus class:

33	42	49	49	53	55	55	61
63	67	68	68	69	69	72	73
74	78	80	83	88	88	88	90
92	94	94	94	94	96	100	

Calculate the following:

- The mean.
- The standard deviation.
- The median.
- The first quartile.
- The third quartile.
- *IQR*.

Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A31.

For the mean:

Function	average
Field 1	A1:A31
Answer	73.5

For the median:

Function	median
Field 1	A1:A31
Answer	73

For the standard deviation:

Function	stdev.s
Field 1	A1:A31
Answer	17.92

For the first quartile:

Function	quartile.exc
Field 1	A1:A31
Field 2	1
Answer	61

For the third quartile:

Function	quartile.exe
Field 1	A1:A31
Field 2	3
Answer	90

For the IQR : $IQR = 90 - 61 = 29$

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and different standard deviations, then comparing the data values directly can be misleading. In order to directly compare values in different data sets, we compare how many standard deviations away the value is from the mean of its data set. This is done by calculating the value's z -score:

Sample	$z = \frac{x - \bar{x}}{s}$
Population	$z = \frac{x - \mu}{\sigma}$

The value x is z standard deviations away from the mean.

EXAMPLE

Two students, John and Ali, are from different high schools and wanted to find out who had the highest GPA when compared to their school. Which student had the highest GPA when compared to their school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Solution

For each student, determine how many standard deviations the z -score is, and their GPA is away from the mean of their school.

$$\text{John: } z = \frac{2.85 - 3.00}{0.7} = -0.21$$

$$\text{Ali: } z = \frac{77 - 80}{10} = -0.3$$

John has a better GPA when compared to his school because his GPA is **0.21** standard deviations **below** his school's mean, while Ali's GPA is **0.3** standard deviations **below** her school's mean. This means that John's GPA is closer to his school's mean than Ali's GPA is to hers.

NOTE

The sign of a z -score is important. A negative z -score tells us that x is below the mean. A positive z -score tells us that x is above the mean. The absolute value of the z -score tells us how many standard deviations the value of x is from the mean.

TRY IT

Two swimmers, Angie and Beth, are from different teams and wanted to find out who had the fastest time for the 50-meter freestyle when compared to her team's mean time. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Click to see Solution

$$\text{Angie: } z = \frac{26.2 - 27.2}{0.8} = -1.25$$

$$\text{Beth: } z = \frac{27.3 - 30.1}{1.4} = -2$$

Angie's time is 1.25 standard deviations **below** her team's mean time, and Beth's is 2 standard deviations **below** her team's time. So, Beth had a faster time when compared to her team's mean than Angie's time is to hers.

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

Chebyshev's Rule: For ANY data set, no matter what the distribution of the data is:

- At least 75\% of the data is within two standard deviations of the mean.
- At least 89\% of the data is within three standard deviations of the mean.
- At least 95\% of the data is within 4.5 standard deviations of the mean.

The Empirical Rule: For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- Approximately 68\% of the data is within one standard deviation of the mean.

- Approximately 95\% of the data is within two standard deviations of the mean.
- More than 99\% of the data is within three standard deviations of the mean.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric.

Exercises

1. How much time does it take to travel to work in a particular region? The table below shows the commute time for a sample of workers in the region who are at least 16 years old and do not work at home.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

- a. Find the range.
- b. Find the standard deviation.
- c. Interpret the standard deviation.
- d. What travel time is one standard deviation above the mean?
- e. What travel time is three standard deviations below the mean?

Click to see Answer

- a. 15.4 minutes
 - b. 3.464 minutes
 - c. On average, the travel time of an arbitrary worker is 3.464 minutes away from the mean 23.462 minutes.
 - d. 26.926 minutes
 - e. 13.07 minutes
2. The following data shows the lengths, in feet, of a sample of boats moored in a marina.

19	35	29	26	21	40	33	33	34
25	20	37	30	26	23	24	29	16
28	25	20	39	32	27	27	27	17

- Find the range.
- Find the standard deviation.
- Interpret the standard deviation.
- What boat length is two standard deviations below the mean?

Click to see Answer

- 24 feet
- 6.48 feet
- On average, an arbitrary boat's length is 6.48 feet away from the mean of 27.33 feet.
- 14.37 feet

3. The data below is the weight, in pounds, of all members of a particular NFL team.

177	210	270	275	212	185	200	241	250
220	259	185	210	272	285	212	250	302
205	232	280	285	184	265	215	223	265
260	278	185	228	273	242	185	241	290
210	276	290	206	174	286	247	190	215
245	205	178	290	280	188	230	260	

- Find the range.
- Find the standard deviation.
- Interpret the standard deviation.
- The team's quarterback weighs 205 pounds. How many standard deviations from the above or below the mean is the quarterback?
- What weight is three standard deviations above the mean?

Click to see Answer

- 128 pounds
- 37.35 pounds
- On average, an arbitrary football player's weight is 37.35 pounds, away from the mean weight of 236.25 pounds.

- d. 0.837 standard deviations below the mean.
- e. 348.3 pounds

4. A sample of 35 post-secondary institutions was taken from across the U.S. The data below shows the number of students enrolled at each institution.

6,414	1,550	2,109	9,350	21,828	4,300	5,944
5,722	2,825	2,044	5,481	5,200	5,853	10,012
6,357	27,000	9,414	7,681	3,200	17,500	9,200
7,380	18,314	6,557	13,713	17,768	7,493	2,771
2,861	1,263	7,285	28,165	5,080	11,622	2,750

- a. Find the range.
- b. Find the standard deviation.
- c. Interpret the standard deviation.
- d. A school with an enrollment of 8,000 students would be how many standard deviations above or below the mean?

Click to see Answer

- a. 26,902 students
 - b. 6,943.89 students
 - c. On average, the number of students enrolled at an arbitrary college is 6,943.89 students, away from the mean of 8,628.74 students.
 - d. 0.09 standard deviations below the mean.
5. Forty randomly selected students were asked the number of pairs of sneakers they owned. The data is recorded below

1	1	2	2	2	2	2	3
3	3	3	3	3	3	3	4
4	4	4	4	4	4	4	4
4	4	4	5	5	5	5	5
5	5	5	5	5	5	5	7

- a. Find the range.

- b. Find the standard deviation.
- c. Interpret the standard deviation.

Click to see Answer

- a. 6 pairs of sneakers
 - b. 1.29 pairs of sneakers
 - c. On average, the number of pairs of sneakers owned by an arbitrary student is 1.29 pairs of sneakers away from the mean of 3.775 pairs of sneakers.
6. Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Click to see Answer

Fredo because his batting average is 0.66 standard deviations below the mean of his team, and Karl's batting average is 0.8 standard deviations below the mean of his team.

7. Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at their school? Explain how you determined your answer.

Student	GPA	School Average GPA	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Click to see Answer

Kamala Thuy's GPA is 0.625 standard deviations below the mean. Vichet's GPA is 0.6 standard deviations above the mean. Kamala's GPA is 1.5 standard deviations above the mean. So Kamala's GPA is the furthest above the mean.

8. A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing \$3,000, a guitar costing \$550, and a drum set costing \$600. The mean cost for a piano is \$4,000 with a standard deviation of \$2,500. The mean cost for a guitar is \$500 with a standard deviation of \$200. The mean cost for drums is \$700 with a standard deviation of \$100. Which cost is the lowest when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type? Justify your answer.

Click to see Answer

Drums. The cost of the drums is 1 standard deviations below the mean. The cost of the piano is 0.4 standard deviations below the mean. The cost of the guitar is 0.25 standard deviations above the mean. So the cost of the drums is the furthest below the mean.

9. An elementary school class ran one mile with a mean of eleven minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran one mile in eight and a half minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.
- Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
 - Who is the fastest runner with respect to his or her class? Explain why.

Click to see Answer

- Kenji's time was 0.25 standard deviations below the mean of their class. Nedda's time was 0.25 standard deviations above the mean of their class. Because Kenji's time was below the mean of their class, Kenji is considered a better run relative to the mean of their class.
- Rachel is the fastest runner because Rachel's time was 1 standard deviation below the mean of their class. This means Rachel's time is the fastest relative to the mean of the class.

10. Using the number of full-time equivalent students (FTES) each year at a local college for the past 40 years, the mean is 1,000 FTES, the median is 1,014 FTES, and the standard deviation is 474 FTES. How many standard deviations above or below the mean is the median?

Click to see Answer

0.0295 standard deviations above the mean.

“2.6 Measures of Dispersion” and “2.7 Exercices” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

PART III

PROBABILITY

It is often necessary to “guess” about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. At a casino where people play games of chance, people select the games based on the belief that the likelihood of winning is good. Students may choose a course of study based on the probable availability of jobs.

Everyone has encountered or used probability at some point in their lives. In fact, most people have an intuitive sense of probability. Probability deals with the chance of an event occurring. In this chapter, we will learn how to solve probability problems using a systematic approach.

CHAPTER OUTLINE

3.1 The Terminology of Probability

3.2 Contingency Tables

3.3 The Complement Rule

3.4 The Addition Rule

3.5 Conditional Probabilities

3.6 Joint Probabilities

“3.1 Introduction to Probability” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

3.1 THE TERMINOLOGY OF PROBABILITY

LEARNING OBJECTIVES

- Define and use the terminology of probability.

Everyday, decisions are made that involve uncertainty about the outcome. The ability to estimate and understand probability helps us make good decisions. **Probability** is a numerical measure that is associated with how certain we are of the outcomes of a particular experiment or activity. Examples of probability used in everyday life include the probability that it will rain today and the probability of winning the lottery.

An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance experiment**. An experiment is any activity where the outcome is uncertain. Flipping a coin, rolling a pair of dice, or drawing a card from a deck of cards are all examples of an experiment.

A result of an experiment is called an **outcome**. For example, in the experiment of flipping a coin, a possible outcome is getting heads. The **sample space** of an experiment is the set of all possible outcomes of that experiment. Three ways to represent a sample space are to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter S is used to denote the sample space. For example, in the experiment of flipping a coin, the sample space has two outcomes: heads or tails. In the notation of probability, we would write the sample space of flipping a coin like $S = \{H, T\}$ where H is heads and T is tails.

An **event** is any combination of outcomes. Generally, an event is a collection of outcomes that possess some trait or characteristic. Upper case letters like A and B are used to represent events. For example, if the experiment is to flip a coin two times, event A might be getting at most one

head in the two flips. In probability, we are interested in finding the probability of an event. The probability of an event A is written $P(A)$.

EXAMPLE

Suppose a coin is flipped two times.

1. What is the sample space for this experiment?
2. Identify all of the outcomes in the event “exactly one head.”
3. Identify all of the outcomes in the event “at least one tail.”

Solution

1. $S = \{HH, HT, TH, TT\}$ where H is heads and T is tails. For example, the outcome HT means heads on the first flip and tails on the second flip.
2. The outcomes in the event “exactly one head” are HT and TH . These are the only outcomes in the sample space S where there is exactly one head in the two flips.
3. The outcomes in the event “at least one tail” are HT , TH , and TT . “At least one” means one or more, so we need to include all of the outcomes in the sample space where there is one or more tails.

NOTE

The order in which things happen is important. The outcomes HT and TH are different outcomes. The outcome HT consists of getting heads on the first flip and tails on the second flip. The outcome TH consists of getting tails on the first flip and heads on the second flip, which is a completely different outcome from HT .

Probability is a numerical measure of the likelihood that an event will occur. The **probability** of an event is the **long-term relative frequency** of that event. Probabilities are numbers between zero and one, inclusive—that is, zero, one, and all numbers between these values. Probabilities

can be written as fractions, decimals, or percents. $P(A) = 0$ means the event A can never happen—the probability is 0%. $P(A) = 1$ means the event A always happens—the probability is 100%. $P(A) = 0.5$ means the event A is equally likely to occur or not to occur—there is a 50% chance A will happen, and a 50% chance A will not happen.

Approaches to Determining Probability

The way that we calculate the probability of an event depends on the situation we are analyzing.

Classical Method Approach to Probability

Most often associated with games of chance, the **classical method approach** requires us to know that the outcomes of an experiment are **equally likely to occur**. We have already seen an experiment where the outcomes are equally likely to occur—flipping a coin. **Equally likely** means that each outcome of an experiment occurs with equal probability. In the experiment of tossing a fair coin, we know that we have a 50% chance of getting heads and a 50% chance of getting tails—the outcomes of heads or tails are equally likely to occur. If we roll a fair, six-sided die, we know that we have the same chance $\left(\frac{1}{6}\right)$ of getting any of the six faces—the outcomes of 1, 2, 3, 4, 5, 6 are equally likely to occur.

To calculate the probability of an event A when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space.

$$P(A) = \frac{\text{number of outcomes in event } A}{\text{total number of outcomes in the sample space}}$$

EXAMPLE

Suppose a coin is flipped two times.

1. What is the probability of getting “exactly one head?”
2. What is the probability of getting “at least one tail?”

Solution

Previously, we found the sample space for this experiment: $S = \{HH, HT, TH, TT\}$.

1. The outcomes in the event “exactly one head” are HT and TH . We see that there are 2 outcomes in the event out of the 4 possible outcomes in the sample space. So

$$P(\text{exactly one head}) = \frac{2}{4} = 0.5$$

2. The outcomes in the event “at least one tail” are HT , TH , and TT . We see that there are 3 outcomes in the event out of the 4 possible outcomes in the sample space. So

$$P(\text{at least one tail}) = \frac{3}{4} = 0.75$$

TRY IT

Suppose we roll a fair six-sided die with the numbers 1, 2, 3, 4, 5, 6 on the faces.

1. What is the sample space for this experiment?

2. What is the probability of getting at least 5?
3. What is the probability of getting an even number?
4. What is the probability of getting a number less than 4?
5. What is the probability of getting a 7?

Click to see Solution

1. $S = \{1, 2, 3, 4, 5, 6\}$
2. $P(\text{at least } 5) = \frac{2}{6} = 0.3333 \dots$
3. $P(\text{even number}) = \frac{3}{6} = 0.5$
4. $P(\text{less than } 4) = \frac{3}{6} = 0.5$
5. $P(7) = \frac{0}{6} = 0$

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair** or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56\% of the time and a tail was obtained 44\% of the time. The data seem to show that the coin is not a fair coin, but more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a board game. The spots on each face are usually small holes carved out and then painted to make the spots visible. The dice may or may not be biased because it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces and the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur.

Empirical Method Approach to Probability

The **empirical** or **relative frequency approach** to probability uses results from identical previous experiments that have been performed many times. Probabilities are based on historical or previously recorded data by determining the proportion of times an event occurs within the data. For example, a retail business owner might want to know the probability that a customer spends

more than \$50 at their store. To determine this probability, the business owner would look at previous sales, count the number of sales over \$50 and then divide that number by the total number of previous sales.

To calculate an empirical probability, repeat the experiment over a large number of trials and record the result of each trial. To find the probability of event A , count the number of times event A happened and divide by the total number of trials.

$$P(A) = \frac{\text{number of times } A \text{ occurs}}{\text{total number of trials}}$$

To get an accurate probability using this approach, it is important that the experiment is repeated a very large number of times. This important characteristic of probability experiments is known as the **law of large numbers**, which states that as the number of repetitions of an experiment increases, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability.

EXAMPLE

An online retailer wants to know the probability that a transaction will be less than \$30. In 2000 transactions, 650 are less than \$30.

Solution

$$P(\text{less than } \$30) = \frac{650}{2000} = 0.325$$

Subjective Method Approach to Probability

In the subjective method approach to probability, probabilities are determined by educated guess, personal belief, intuition, or expert reasoning. A subjective probability is essentially a guess, but a guess based on an accumulation of knowledge, understanding, and experience. Estimating the

probability the price of a stock goes down over time or the probability a certain sports team will win a championship are examples of subjective probability.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=85#oembed-1>

Video: “Probability: Tossing 2 Coins (Head/Tail)” by Joshua Emmanuel [5:56] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A box is filled with several party favours. It contains 12 hats, 15 noisemakers, 10 finger traps, and 5 bags of confetti. A party guest randomly selects one of the party favours from the box.
 - a. Find the probability of getting a hat.
 - b. Find the probability of getting a noisemaker.
 - c. Find the probability of getting a finger trap.
 - d. Find the probability of getting a bag of confetti.

Click to see Answer

- a. 0.2857
 - b. 0.3571
 - c. 0.2381
 - d. 0.119

2. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange. A jelly bean is selected at random.
 - a. Find the probability of getting a blue jelly bean.
 - b. Find the probability of getting a green jelly bean.
 - c. Find the probability of getting a purple jelly bean.
 - d. Find the probability of getting a red jelly bean.

- e. Find the probability of getting a yellow jelly bean.
- f. Find the probability of getting an orange jelly bean.

Click to see Answer

- a. 0.1733
- b. 0.1333
- c. 0.1867
- d. 0.1467
- e. 0.2533
- f. 0.1067

3. Suppose a card is drawn from a standard deck of 52 cards.
- a. What is the probability the card is red?
 - b. What is the probability the card is a club?
 - c. What is the probability the card is a face card (jack, queen, or king)?
 - d. What is the probability the card is an ace?
 - e. What is the probability the card is the jack of hearts?

Click to see Answer

- a. 0.5
- b. 0.25
- c. 0.2308
- d. 0.0769
- e. 0.0192

4. Suppose a fair, six-sided die is rolled.
- a. What is the probability of rolling an even number?
 - b. What is the probability of rolling a 2, 3, or 5?
 - c. What is the probability of rolling a 3 or a 6?

Click to see Answer

- a. 0.5
- b. 0.5
- c. 0.3333

5. What is the word for the set of all possible outcomes?

Click to see Answer

sample space.

6. A sample of students was surveyed and asked how many movies they watched last week. The results are given in the table below.

Number of Movies	Frequency
0	5
1	9
2	6
3	4
4	1

- What is the probability a student watched 0 movies last week?
- What is the probability a student watched at most 1 movie last week?
- What is the probability a student watched 2 or more movies last week?

Click to see Answer

- 0.2
- 0.56
- 0.44

7. A sample of students was surveyed and asked how many pairs of sneakers they own. The results are given in the table below.

Number of Pairs of Sneakers	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

- What is the probability a student owns exactly 2 pairs of sneakers?

- b. What is the probability a student owns exactly 6 pairs of sneakers?
- c. What is the probability a student owns 4 or more pairs of sneakers?
- d. What is the probability a student owns 3 or fewer pairs of sneakers?
- e. What is the probability a student owns 5 or 7 pairs of sneakers?

Click to see Answer

- a. 0.125
- b. 0
- c. 0.625
- d. 0.375
- e. 0.325

“3.2 The Terminology of Probability” and “3.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

3.2 CONTINGENCY TABLES

LEARNING OBJECTIVES

- Construct and interpret contingency tables.
- Find probabilities using contingency tables.

A **contingency table** provides a way of displaying data that can facilitate calculating probabilities. The table can be used to describe the sample space of an experiment. Contingency tables allow us to break down a sample space when two variables are involved.

Cell Phone Use	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

When reading a contingency table:

- The left-side column lists all of the values for one of the variables. In the table shown above, the left-side column shows the variable about whether or not someone uses a cell phone while driving.
- The top row lists all of the values for the other variable. In the table shown above, the top row shows the variable about whether or not someone had a speeding violation in the last year.
- In the body of the table, the cells contain the number of outcomes that fall into both of the categories corresponding to the intersecting row and column. In the table shown above, the number of 25 at the intersection of the “cell phone user” row and “speeding violation in the

last year” column tells us that there are 25 people who have both of these characteristics.

- The bottom row gives the totals in each column. In the table shown above, the number 685 in the bottom of the “no speeding violation in the last year” tells us that there are 685 people who did not have a speeding violation in the last year.
- The right-side column gives the totals in each row. In the table shown above, the number 305 in the right side of the “cell phone user” row tells us that there are 305 people who use cell phones while driving.
- The number in the bottom right corner is the size of the sample space. In the table shown above, the number in the bottom right corner is 755, which tells us that there are 755 people in the sample space.

EXAMPLE

Suppose a study of speeding violations and drivers who use cell phones while driving produced the following fictional data:

Cell Phone Use	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

Calculate the following probabilities:

1. What is the probability that a randomly selected person is a cell phone user?
2. What is the probability that a randomly selected person had no speeding violations in the last year?
3. What is the probability that a randomly selected person had a speeding violation in the last year and does not use a cell phone?
4. What is the probability that a randomly selected person uses a cell phone and had no speeding violations in the last year?

Solution

1. Probability = $\frac{\text{number of cell phone users}}{\text{total number in study}} = \frac{305}{755}$
2. Probability = $\frac{\text{number of no violations}}{\text{total number in study}} = \frac{685}{755}$
3. Probability = $\frac{\text{number of violations and not cell phone users}}{\text{total number in study}} = \frac{45}{755}$
4. Probability = $\frac{\text{number of cell phone users and no violations}}{\text{total number in study}} = \frac{280}{755}$

TRY IT

The table below shows the number of athletes who stretch before exercising and how many had injuries within the past year.

Stretch habits	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

1. What is the probability that a randomly selected athlete stretches before exercising?
2. What is the probability that a randomly selected athlete had an injury in the last year?
3. What is the probability that a randomly selected athlete does not stretch before exercising and had no injuries in the last year?
4. What is the probability that a randomly selected athlete stretches before exercising and had no injuries in the last year?

Click to see Solution

1. Probability = $\frac{350}{800} = 0.4375$

2. Probability = $\frac{286}{800} = 0.3575$
3. Probability = $\frac{219}{800} = 0.27375$
4. Probability = $\frac{295}{800} = 0.36875$

EXAMPLE

The table below shows a random sample of 100 hikers broken down by gender and the areas of hiking they prefer.

Gender	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16		45
Male			14	55
Total		41		

1. Fill in the missing values in the table
2. What is the probability that a randomly selected hiker is female?
3. What is the probability that a randomly selected hiker prefers to hike on the coast?
4. What is the probability that a randomly selected hiker is male and prefers to hike near lakes and streams?
5. What is the probability that a randomly selected hiker is female and prefers to hike in the mountains?

Solution

1.

Gender	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	11	45
Male	16	25	14	55
Total	34	41	25	100

2. Probability = $\frac{45}{100} = 0.45$
3. Probability = $\frac{34}{100} = 0.34$
4. Probability = $\frac{25}{100} = 0.25$
5. Probability = $\frac{11}{100} = 0.11$

TRY IT

The table below relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

- Find the total for each row and column.
- Find the probability that a randomly chosen individual from this group is tall.
- Find the probability that a randomly chosen individual from this group is normal.
- Find the probability that a randomly chosen individual from this group is obese and short.

5. Find the probability that a randomly chosen individual from this group is underweight and medium.

Click to see Solution

1.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	60
Normal	20	51	28	99
Underweight	12	25	9	46
Totals	50	104	51	205

2. Probability = $\frac{50}{205}$
3. Probability = $\frac{99}{205}$
4. Probability = $\frac{14}{205}$
5. Probability = $\frac{25}{205}$



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=87#oembed-1>

Video: “Ex: Basic Example of Finding Probability From a Table” by Mathispower4u [2:40] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A previous year, the weights, in pounds, of the members of the **San Francisco 49ers** and the

Dallas Cowboys were published in the SAN JOSE MERCURY NEWS. The factual data are compiled into the table.

Shirt Number	At most 210	211–250	251–290	More than 290	Total
1–33	21	5	0	0	26
34–66	6	18	7	4	35
67–99	6	12	22	5	45
Total	33	35	29	9	106

For the following, suppose that one player from these two teams is selected at random.

- What is the probability that the player's shirt number is in the 34-66 category?
- What is the probability that the player weighs at most 210 pounds?
- What is the probability that the player's shirt number is in the 1-33 category and weighs between 211 and 250 pounds?

Click to see Answer

- 0.3302
- 0.3113
- 0.0472

- The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home Run	Total Hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

- What is the probability that a hit was made by Jackie Robinson?
- What is the probability that the hit was a double?
- What is the probability that the hit was made by Ty Cobb?
- What is the probability that a hit was made by Hank Aaron and is a home run?
- What is the probability that a hit was a triple and hit by Babe Ruth?

Click to see Answer

- a. 0.1229
- b. 0.1277
- c. 0.3392
- d. 0.0611
- e. 0.011

3. The table shows a random sample of musicians and how they learned to play their instruments.

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

- a. Find the probability a musician is female.
- b. Find the probability that a musician received private instruction.
- c. Find the probability that a musician is male and is self-taught.

Click to see Answer

- a. 0.5538
- b. 0.2846
- c. 0.1462

4. The table shows the political party affiliation of each of 67 members of the US Senate in June 2012, and when they are up for reelection.

Up for reelection:	Democratic Party	Republican Party	Other	Total
November 2014	20	13	0	
November 2016	10	24	0	
Total				

- a. Complete the table by filling in the totals.
- b. What is the probability that a randomly selected senator has an “Other” affiliation?
- c. What is the probability that a randomly selected senator is a Republican?

- d. What is the probability that a randomly selected senator is up for reelection in November 2016?
- e. What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2014?

Click to see Answer

a.

Up for reelection:	Democratic Party	Republican Party	Other	Total
November 2014	20	13	0	33
November 2016	10	24	0	34
Total	30	37	0	67

- b. 0
- c. 0.5522
- d. 0.5075
- e. 0.2985

5. The table below identifies a group of children by one of four hair colours and by type of hair.

Hair Type	Brown	Blond	Black	Red	Totals
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

- a. Complete the table.
- b. What is the probability that a randomly selected child will have wavy hair?
- c. What is the probability that a randomly selected child will have blond hair?
- d. What is the probability that a randomly selected child will have straight hair and brown hair?

Click to see Answer

a.

Hair Type	Brown	Blond	Black	Red	Totals
Wavy	20	5	15	3	43
Straight	80	15	65	12	172
Totals	100	20	80	15	215

- b. 0.2
- c. 0.093
- d. 0.3721

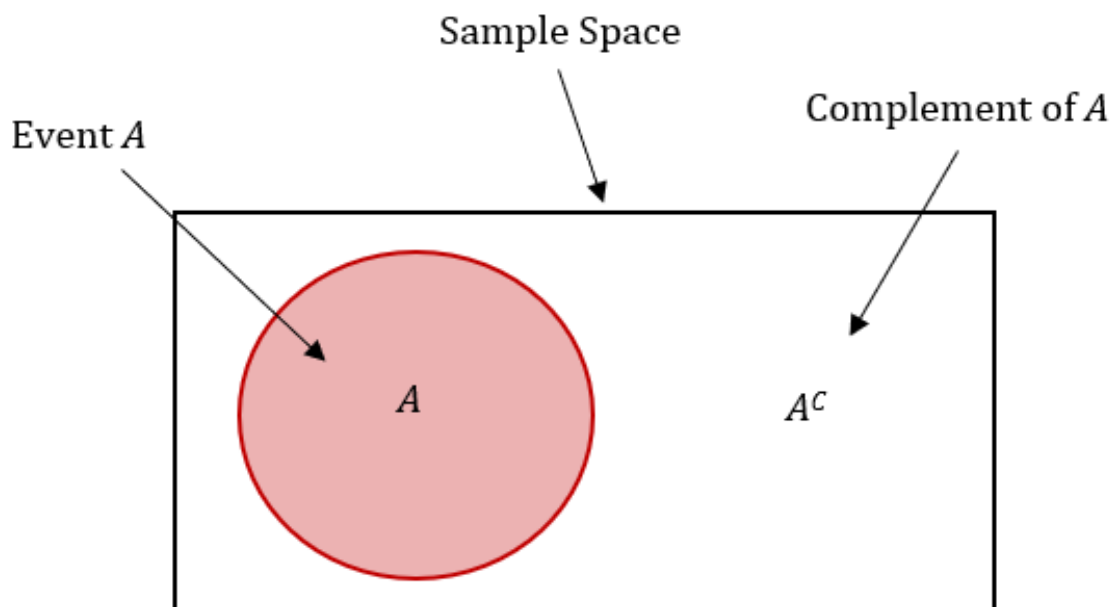
“3.3 Contingency Tables” and “3.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

3.3 THE COMPLEMENT RULE

LEARNING OBJECTIVES

- Calculate probabilities using the complement rule.

The **complement** of an event A is the set of all outcomes in the sample space that are not in A . The complement of A is denoted by A^C and is read “not A .”



EXAMPLE

Suppose a coin is flipped two times. Previously, we found the sample space for this experiment: $S = \{HH, HT, TH, TT\}$ where H is heads and T is tails.

1. What is the complement of the event “exactly one head”?
2. What is the complement of the event “at least one tail.”

Solution

1. The event “exactly one head” consists of the outcomes HT and TH . The **complement** of “exactly one head” consists of the outcomes HH and TT . These are the outcomes in the sample space S that are NOT in the original event “exactly one head.”
2. The event “at least one tail” consists of the outcomes HT , TH , and TT . The **complement** of “at least one tail” consists of the outcomes HH . These are the outcomes in the sample space S that are NOT in the original event “at least one tail.”

TRY IT

Suppose we roll a fair six-sided die with the numbers 1, 2, 3, 4, 5, 6 on the faces. Previously, we found the sample space for this experiment: $S = \{1, 2, 3, 4, 5, 6\}$

1. What is the complement of the event “rolling a 4”?
2. What is the complement of the event “rolling a number greater than or equal to 5”?
3. What is the complement of the event “rolling an even number”?
4. What is the complement of the event “rolling a number less than 4”?

Click to see Solution

1. The complement is $\{1, 2, 3, 5, 6\}$.
2. The complement is $\{1, 2, 3, 4\}$.
3. The complement is $\{1, 3, 5\}$.
4. The complement is $\{4, 5, 6\}$.

The Probability of the Complement

In any experiment, an event A or its complement A^C must occur. This means that $P(A) + P(A^C) = 1$. Rearranging this equation gives us a formula for finding the probability of the complement from the original event:

$$P(A^C) = 1 - P(A)$$

EXAMPLE

An online retailer knows that 30% of customers spend more than \$100 per transaction. What is the probability that a customer spends at most \$100 per transaction?

Solution

Spending at most \$100 (\$100 or less) per transaction is the complement of spending more than \$100 per transaction.

$$\begin{aligned}
 P(\text{at most } \$100) &= 1 - P(\text{more than } \$100) \\
 &= 1 - 0.3 \\
 &= 0.7
 \end{aligned}$$

TRY IT

At a local college, a statistics professor has a class of 80 students. After polling the students in the class, the professor finds out that 15 of the students play on one of the school's sports teams and 60 of the students have part-time jobs.

1. What is the probability that a student in the class does not play on one of the school's sports teams?
2. What is the probability that a student in the class does not have a part-time job?

Click to see Solution

1. $P(\text{no sports team}) = 1 - P(\text{sports team}) = 1 - \frac{15}{80} = 0.8125$
2. $P(\text{no part-time job}) = 1 - P(\text{part-time job}) = 1 - \frac{60}{80} = 0.25$

Exercises

1. Suppose a coin is flipped three times.
 - a. Find the sample space for this experiment.
 - b. What is the complement of the event “exactly 1 tail”?
 - c. What is the complement of the event “at most 2 heads”?
 - d. What is the complement of the event “2 or more tails”?

Click to see Answer

- a. $\{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$
- b. $\{HHH, TTH, THT, HTT, TTT\}$
- c. $\{HHH\}$
- d. $\{HHH, HHT, HTH, THH\}$

2. A 12-sided die is in the shape of a regular dodecahedron. The faces of the 12-sided die are labelled with the numbers 1 to 12. Suppose the 12-sided die is rolled one time.
- Find the sample space of this experiment.
 - What is the complement of the event “rolling a 7”?
 - What is the complement of the event “rolling a number less than or equal to 9”?
 - What is the complement of the event “rolling a number that is a multiple of 3”?
 - What is the complement of the event “rolling a 5 or 9 or 12”?
 - What is the complement of the event “rolling a number greater than 8”?
 - What is the complement of the event “rolling an odd number”?

Click to see Answer

- $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$
 - $\{1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12\}$
 - $\{10, 11, 12\}$
 - $\{1, 2, 4, 5, 7, 8, 10, 11\}$
 - $\{1, 2, 3, 4, 6, 7, 8, 10, 11\}$
 - $\{1, 2, 3, 4, 5, 6, 7, 8\}$
 - $\{2, 4, 6, 8, 10, 12\}$
3. A recent survey asked people about home ownership and annual income. A total of 750 people were surveyed. Of the 750 people surveyed, 425 owned a home. Of the 750 people surveyed, 338 people had an annual income of \$60,000 or more.
- What is the probability that one of the people in the survey does not own a home?
 - What is the probability that one of the people in the survey has an annual income of less than \$60,000?

Click to see Answer

- 0.4333
 - 0.5493
4. A local college surveyed its recent graduates about their overall satisfaction with their college experience and employment status post-graduation. In the survey, 75% of respondents said they were satisfied with their college experience, and 64% of respondents said they found full-time jobs after graduation.
- What is the probability that a respondent was not satisfied with their college experience?
 - What is the probability that a respondent did not find full-time jobs after graduation?

Click to see Answer

- 0.25

b. 0.36

“3.4 The Complement Rule” and “3.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

3.4 THE ADDITION RULE

LEARNING OBJECTIVES

- Calculate “or” probabilities using the addition rule.
- Determine if two events are mutually exclusive.

For two events A and B we might want to know the probability that at least one of the two events occurs. For example, we might want to find the probability of rolling a 2 or a 5 in a single roll of a die, or we might want to find the probability that someone has a smartphone or a tablet. In probability terms, we want to find $P(A \text{ or } B)$, the probability that either A or B occurs. In probability, “or” is always an **inclusive** “or,” which means that either A occurs, or B occurs, or both occur.

The Addition Rule for Probabilities

To find $P(A \text{ or } B)$, we start by adding the individual probabilities, $P(A)$ and $P(B)$. But this means that the overlap between the two events A and B is counted **twice**: once by $P(A)$ and once by $P(B)$. To correct for this double counting, we need to subtract $P(A \text{ and } B)$, the probability of both events occurring. This gives us the addition rule to find $P(A \text{ or } B)$:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

EXAMPLE

At a local language school, 40\% of the students are learning Spanish, 20\% of the students are learning German, and 8\% of the students are learning both Spanish and German. What is the probability that a randomly selected student is learning Spanish or German?

Solution

$$\begin{aligned} P(\text{Spanish or German}) &= P(\text{Spanish}) + P(\text{German}) - P(\text{Spanish and German}) \\ &= 0.4 + 0.2 - 0.08 \\ &= 0.52 \end{aligned}$$

EXAMPLE

There are 50 students enrolled in the second year of a business degree program. During this semester, the students have to take some elective courses. 18 students decide to take an elective in psychology, 27 students decide to take an elective in philosophy, and 10 students decide to take an elective in both psychology and philosophy. What is the probability that a student takes an elective in psychology or philosophy?

Solution

$$\begin{aligned} P(\text{psychology or philosophy}) &= P(\text{psychology}) + P(\text{philosophy}) \\ &\quad - P(\text{psychology and philosophy}) \\ &= \frac{18}{50} + \frac{27}{50} - \frac{10}{50} \\ &= 0.7 \end{aligned}$$

TRY IT

At a local basketball game, 70\% of the fans are cheering for the home team, 25\% of the fans are wearing blue, and 12\% of the fans are cheering for the home team and wearing blue. What is the probability that a randomly selected fan is cheering for the home team or wearing blue?

Click to see Solution

$$\begin{aligned} P(\text{home team or blue}) &= P(\text{home team}) + P(\text{blue}) - P(\text{home team and blue}) \\ &= 0.7 + 0.25 - 0.12 \\ &= 0.83 \end{aligned}$$

EXAMPLE

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

Cell phone use	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

1. What is the probability that a randomly selected person is a cell phone user or has no speeding violations in the last year?

2. What is the probability that a randomly selected person had a speeding violation in the last year or does not use a cell phone?

Solution

$$P(\text{cell phone or no violations}) = P(\text{cell phone}) + P(\text{no violations}) - P(\text{cell phone and no violations})$$

$$\begin{aligned} 1. \qquad &= \frac{305}{755} + \frac{685}{755} - \frac{280}{755} \\ &= \frac{710}{755} \end{aligned}$$

$$P(\text{violations or no cell phone}) = P(\text{violations}) + P(\text{no cell phone}) - P(\text{violations and no cell phone})$$

$$\begin{aligned} 2. \qquad &= \frac{70}{755} + \frac{450}{755} - \frac{45}{755} \\ &= \frac{475}{755} \end{aligned}$$

TRY IT

This table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

Stretching Practice	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

1. What is the probability that a randomly selected athlete stretches before exercising or had an injury last year?
2. What is the probability that a randomly selected athlete does not stretch before exercising or had no injuries in the last year?

Click to see Solution

$$1. \text{ Probability} = \frac{350}{800} + \frac{286}{800} - \frac{55}{800} = 0.72625$$

$$2. \text{ Probability} = \frac{450}{800} + \frac{514}{800} - \frac{219}{800} = 0.93125$$

Mutually Exclusive Events

Two events, A and B , are **mutually exclusive** if the two events cannot happen at the same time. That is, the events A and B do not share any outcomes, and so $P(A \text{ and } B) = 0$. For example, in the experiment of flipping a coin, the events heads and tails are mutually exclusive because it is not possible to have both heads and tails on the top face. In the case of mutually exclusive events, the addition rule is $P(A \text{ or } B) = P(A) + P(B)$.

EXAMPLE

Suppose a bag contains 20 balls. 10 of the balls are white, 7 of the balls are red, and 3 of the balls are blue. Suppose one ball is selected at random from the bag.

1. Are the events “selecting a white ball” and “selecting a red ball” mutually exclusive? Why?
2. What is the probability of selecting a white or red ball?

Solution

1. The events “selecting a white ball” and “selecting a red ball” are mutually exclusive because the events cannot happen at the same time. It is not possible for the selected ball to be both white and red.

$$2. P(\text{white or red}) = P(\text{white}) + P(\text{red}) = \frac{10}{20} + \frac{7}{20} = 0.85$$

NOTE

In the calculation of the probability in part 2, there is nothing to subtract. Because the events are mutually exclusive, $P(\text{white and red}) = 0$.

TRY IT

At a local college, 60% of the students are taking a math class, 50% of the students are taking a science class, and 30% of the students are taking both a math and a science class.

1. Are the events “taking a math class” and “taking a science class” mutually exclusive? Explain.
2. What is the probability that a randomly selected student is taking a math class or a science class?

Click to see Solution

1. The events “taking a math class” and “taking a science class” are not mutually exclusive because the events can happen at the same time (i.e. a student can be taking both a math class and a science class). As stated in the question, $P(\text{math and science}) = 0.3 \neq 0$.
2. $P(\text{math or science}) = P(\text{math}) + P(\text{science}) - P(\text{math and science}) = 0.6 + 0.5 - 0.3 = 0.8$

TRY IT

A fair, six-sided die is rolled one time.

1. Are the events “rolling a 4” and “rolling an even number” mutually exclusive?
2. Are the events “rolling a 4” and “rolling an odd number” mutually exclusive?
3. What is the probability of rolling a 4 or rolling an odd number?

Click to see Solution

1. The events “rolling a 4” and “rolling an even number” are not mutually exclusive because the events can happen at the same time (i.e. 4 is an even number).
2. The events “rolling a 4” and “rolling an odd number” are mutually exclusive because the events cannot happen at the same time. It is not possible to roll a die and get a 4 (an even number) and an odd number on the top face at the same time
3. $P(4 \text{ or odd}) = P(4) + P(\text{odd}) = \frac{1}{6} + \frac{3}{6} = \frac{4}{6}$



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=92#oembed-1>

Video: “Addition rule for probability | Probability and Statistics | Khan Academy” by Khan Academy [10:43] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A 12-sided die is in the shape of a regular dodecahedron. The faces of the 12-sided die are labelled with the numbers 1 to 12. Suppose the 12-sided die is rolled one time.
 - a. Are the events “rolling an even number” and “rolling a multiple of 3” mutually exclusive? Explain
 - b. What is the probability of “rolling an even number” or “rolling a multiple of 3”?
 - c. Are the events “rolling a number greater than 9” and “rolling a number less than 5” mutually exclusive? Explain.
 - d. What is the probability of “rolling a number greater than 9” or “rolling a number less than 5”?

Click to see Answer

- a. Not mutually exclusive because the events can happen at the same time (i.e. 6 is an even number and is a multiple of 3).
 - b. 0.6667
 - c. Mutually exclusive because it is not possible to get a number greater than 9 (i.e. 10, 11, 12) and a number less than 5 (i.e. 1, 2, 3, 4) at the same time.
 - d. 0.5833
2. A recent survey asked people about home ownership and annual income. A total of 750 people were surveyed. Of the 750 people surveyed, 425 owned a home, 338 people had an annual income of \$60,000 or more, and 293 people owned a home and had an annual income of \$60,000 or more.
 - a. Are the events “owned a home” and “annual income of \$60,000 or more” mutually

exclusive? Explain.

- b. What is the probability that one of the people in the survey owned a home or had an annual income of \$60,000 or more?

Click to see Answer

- a. Not mutually exclusive because 293 people fall into both categories.
- b. 0.6267

3. A local college surveyed its recent graduates about their overall satisfaction with their college experience and employment status post-graduation. In the survey, 75% of respondents said they were satisfied with their college experience, 64% of respondents said they found full-time jobs after graduation, and 52% of respondents said they were satisfied with their college experience and found full-time jobs after graduation.

- a. Are the events “satisfied with college experience” and “found full-time job” mutually exclusive? Explain.
- b. What is the probability that a respondent was satisfied with their college experience and found full-time jobs after graduation?

Click to see Answer

- a. Not mutually exclusive because 52% fall into both categories.
- b. 0.87

4. U and V are mutually exclusive events. $P(U) = 0.26$; $P(V) = 0.37$. Find:

- a. $P(U \text{ and } V)$
- b. $P(U \text{ or } V)$

Click to see Answer

- a. 0
- b. 0.63

5. At a local college, 20% of the students are studying business, 40% of the students are studying mathematics, and 8% of the students are studying both business and mathematics.

- a. What is the probability that a randomly selected student studies business or mathematics?
- b. Are the events “business” and “mathematics” mutually exclusive? Explain.

Click to see Answer

- a. 0.52
- b. Not mutually exclusive because 8% of the students are studying both.

6. In a collection of eight cards, five cards are green, and three cards are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. One card is selected at random.
- What is the probability the card is green?
 - What is the probability the card is green or has an even number on it?
 - What is the probability the card is green and has an even number on it?
 - Are the events “green” and “even” mutually exclusive? Explain.
 - What is the probability the card is yellow or has a number greater than 3 on it?
 - Are the events “yellow” and “number greater than 3” mutually exclusive? Explain.

Click to see Answer

- 0.625
 - 0.75
 - 0.25
 - Not mutually exclusive because there are green cards with even numbers.
 - 0.625
 - Mutually exclusive because there are no yellow cards with numbers greater than 3.
7. Canadian Blood Services collects blood donations. A person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Data shows that 43\% of people have type O blood, 15\% of people have Rh- factor and 52\% of people have type O or Rh- factor.
- Find the probability that a person has both type O blood and the Rh- factor.
 - Find the probability that a person does NOT have both type O blood and the Rh- factor.

Click to see Answer

- 0.06
 - 0.48
8. At a college, 72\% of courses have final exams, 46\% of courses require research papers, and 32\% of courses have both a research paper and a final exam.
- Find the probability that a course has a final exam or a research project.
 - Find the probability that a course has NEITHER of these two requirements.

Click to see Answer

- 0.86
- 0.14

9. In a box of assorted cookies, 36\% contain chocolate, 12\% contain nuts, and 8\% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.
- Find the probability that a cookie contains chocolate or nuts (Sean can't eat it).
 - Find the probability that a cookie does not contain chocolate or nuts (Sean can eat it).

Click to see Answer

- 0.4
- 0.6

10. A previous year, the weights, in pounds, of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the SAN JOSE MERCURY NEWS. The factual data are compiled into the table.

Shirt Number	At most 210	211–250	251–290	More than 290	Total
1–33	21	5	0	0	26
34–66	6	18	7	4	35
67–99	6	12	22	5	45
Total	33	35	29	9	106

For the following, suppose that one player from these two teams is selected at random.

- What is the probability that the player's weighs at most 210 pounds or has a shirt number in the 67-99 category?
- What is the probability that the player weighs between 251 and 290 pounds or has a shirt number in the 34-66 category?
- Are the events "1-33" and "more than 290" mutually exclusive? Explain

Click to see Answer

- 0.6792
- 0.5377
- Mutually exclusive because there are no players that fall into both of those categories.

11. The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home Run	Total Hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

- What is the probability that a hit was made by Jackie Robinson or was a double?
- What is the probability that the hit was a single or made by Hank Aaron?
- Are the events “Babe Ruth” and “home run” mutually exclusive? Explain.

Click to see Answer

- 0.2285
- 0.8054
- Not mutually exclusive because Babe Ruth hit 714 home runs.

12. The table shows a random sample of musicians and how they learned to play their instruments.

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

- Find the probability a musician is male or is self-taught.
- Find the probability that a musician studied in school or is female.

Click to see Answer

- 0.5385
- 0.7385

13. The table shows the political party affiliation of each of 67 members of the US Senate in June 2012 and when they are up for reelection.

Up for reelection:	Democratic Party	Republican Party	Other	Total
November 2014	20	13	0	33
November 2016	10	24	0	34
Total	30	37	0	67

- What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2016?
- What is the probability that a randomly selected senator is a Democrat or is up for reelection in November 2014?

Click to see Answer

- 0.7015
- 0.6418

14. The table below identifies a group of children by one of four hair colours and by type of hair.

Hair Type	Brown	Blond	Black	Red	Totals
Wavy	20	5	15	3	43
Straight	80	15	65	12	172
Totals	100	20	80	15	215

- What is the probability that a randomly selected child has wavy hair or black hair?
- What is the probability that a randomly selected child has blond hair or straight hair?
- Are the events “wavy” or “brown” mutually exclusive? Explain.

Click to see Answer

- 0.5023
- 0.8233
- Not mutually exclusive because 20 children fall into both categories.

3.5 CONDITIONAL PROBABILITY

LEARNING OBJECTIVES

- Calculate conditional probabilities.
- Determine if two events are independent.

A **conditional probability** is the probability of an event A **given** that another event B has already occurred. The idea behind conditional probability is that it reduces the sample space to the part of the sample space that involves just the given event B —except for the event B , everything else in the sample space is thrown away. Once the sample space is reduced to the given event B , we calculate the probability of A occurring within the reduced sample space.

The conditional probability of A given B is written as $P(A|B)$ and is read as “the probability of A given B .”

Recognizing a conditional probability and identifying which event is the given event can be challenging. The following sentences are all asking the same conditional probability, just in different ways:

- What is the probability a student has a smartphone, given that the student has a tablet?
- If a student has a tablet, what is the probability the student has a smartphone?
- What is the probability that a student with a tablet has a smartphone?

The given event is “has a tablet,” so in calculating the conditional probability, we would restrict the sample space to just those students who have a tablet and then find the probability a student has a smartphone from among just those students with a tablet.

NOTE

The conditional probability $P(A|B)$ is **NOT** the same as $P(A \text{ and } B)$.

- In the conditional probability $P(A|B)$, we want to find the probability of A occurring **after** B has already happened. In the conditional probability, the sample space is restricted to just event B before we calculate the probability of A in the restricted sample space.
- In $P(A \text{ and } B)$, we want to find the probability of events A and B happening **at the same time** in the unrestricted sample space.

EXAMPLE

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

Cell Phone Use	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

1. What is the probability that a randomly selected person is a cell phone user, given that they had no speeding violations in the last year?
2. If a randomly selected person does not have a cell phone, what is the probability they had a speeding violation last year?
3. What is the probability that someone with a cell phone did not have a speeding violation last year?

Solution

1. The given event is "no speeding violations," so we restrict the table to just the column involving "no speeding violations." With this restriction, the table would look like this:

Cell Phone Use	No speeding violation in the last year
Cell phone user	280
Not a cell phone user	405
Total	685

Now, we want to find the probability a person is a cell phone user in this restricted sample space:

$$\begin{aligned}
 P(\text{cell phone}|\text{no violations}) &= \frac{\text{number of cell phone users in restricted sample space}}{\text{total number in restricted sample space}} \\
 &= \frac{280}{685}
 \end{aligned}$$

2. The given event is "no cell phone," so we restrict the table to just the row involving "no cell phone." With this restriction, the table would look like this:

Cell Phone Use	Speeding violation in the last year	No speeding violation in the last year	Total
Not a cell phone user	45	405	450

Now, we want to find the probability a person has a speeding violation in the last year in this restricted sample space:

$$\begin{aligned}
 P(\text{violation}|\text{no cell phone}) &= \frac{\text{number of violations in restricted sample space}}{\text{total number in restricted sample space}} \\
 &= \frac{45}{450}
 \end{aligned}$$

3. The given event is "cell phone," so we restrict the table to just the row involving "cell phone." With this restriction, the table would look like this:

Cell Phone Use	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305

Now, we want to find the probability a person does not have a speeding violation in the last year in this restricted sample space:

$$\begin{aligned}
 P(\text{no violations}|\text{cell phone}) &= \frac{\text{number with no violations in restricted sample space}}{\text{total number in restricted sample space}} \\
 &= \frac{280}{305}
 \end{aligned}$$

NOTE

The conditional probability $P(A|B)$ does not equal the conditional probability $P(B|A)$. In

the above example, $P(\text{cell phone}|\text{no violations}) = \frac{280}{685}$ **does not equal**

$$P(\text{no violations}|\text{cell phone}) = \frac{280}{305}.$$

TRY IT

This table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

Stretching Practice	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

1. What is the probability that a randomly selected athlete stretches before exercising, given that they had an injury last year?
2. What is the probability that a randomly selected athlete who had no injuries in the last year does not stretch before exercising?
3. If a randomly selected athlete does not stretch before exercising, what is the probability they had an injury in the last year?

Click to see Solution

$$1. \text{ Probability} = \frac{55}{286}$$

$$2. \text{ Probability} = \frac{219}{514}$$

$$3. \text{ Probability} = \frac{231}{450}$$

Calculating Conditional Probabilities Using the Formula

When working with a contingency table as in the above examples, we can simply calculate conditional probabilities by restricting the table to the given event and then finding the required probability in the restricted sample space. Depending on the situation, it might not be possible to work out a conditional probability this way. In these situations, we can use the following formula to find a conditional probability:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

EXAMPLE

At a local language school, 40\% of the students are learning Spanish, 20\% of the students are learning German, and 8\% of the students are learning both Spanish and German.

1. What is the probability that a randomly selected student is learning Spanish given that they are learning German?
2. What is the probability that a randomly selected Spanish student is learning German?

Solution

$$1. P(\text{Spanish}|\text{German}) = \frac{P(\text{Spanish and German})}{P(\text{German})} = \frac{0.08}{0.2} = 0.4$$

$$2. P(\text{German}|\text{Spanish}) = \frac{P(\text{Spanish and German})}{P(\text{Spanish})} = \frac{0.08}{0.4} = 0.2$$

EXAMPLE

There are 50 students enrolled in the second year of a business degree program. During this semester, the students have to take some elective courses. 18 students decide to take an elective in psychology, 27 students decide to take an elective in philosophy, and 10 students decide to take an elective in both psychology and philosophy.

1. What is the probability that a student takes an elective in psychology given that they take an elective in philosophy?
2. If a student takes an elective in psychology, what is the probability that they take an elective in

philosophy?

Solution

$$1. \ P(\text{psychology}|\text{philosophy}) = \frac{P(\text{psychology and philosophy})}{P(\text{philosophy})} = \frac{\frac{10}{50}}{\frac{27}{50}} = 0.3704$$

$$2. \ P(\text{philosophy}|\text{psychology}) = \frac{P(\text{psychology and philosophy})}{P(\text{psychology})} = \frac{\frac{10}{50}}{\frac{18}{50}} = 0.5556$$

TRY IT

At a local basketball game, 70% of the fans are cheering for the home team, 25% of the fans are wearing blue, and 12% of the fans are cheering for the home team and wearing blue.

1. What is the probability that a randomly selected fan is cheering for the home team, given that they are wearing blue?
2. If a randomly selected fan is cheering for the home team, what is the probability they are wearing blue?

Click to see Solution

$$1. \ P(\text{home team}|\text{blue}) = \frac{0.12}{0.25} = 0.48$$

$$2. \ P(\text{blue}|\text{home team}) = \frac{0.12}{0.7} = 0.1714$$

Independent Events

Two events are **independent** if the probability of the occurrence of one of the events does not affect the probability of the occurrence of the other event. In other words, two events, A and B are independent if the knowledge that one of the events occurred does not affect the chance the other event occurs. For example, the outcomes of two rolls of a fair die are independent events—the outcome of the first roll does not change the probability of the outcome of the second roll. If two events are not independent, then we say the events are **dependent**.

We can test two events A and B for independence by comparing $P(A)$ and $P(A|B)$:

- If $P(A) = P(A|B)$, then the events A and B are independent.
- If $P(A) \neq P(A|B)$, then the events A and B are dependent.

EXAMPLE

At a local language school, 40% of the students are learning Spanish, 20% of the students are learning German, and 8% of the students are learning both Spanish and German. Are the events "Spanish" and "German" independent? Explain.

Solution

To check for independence, we need to check two probabilities: $P(\text{Spanish})$ and $P(\text{Spanish}|\text{German})$. If these probabilities are equal, the events are independent. If the probabilities are not equal, the events are dependent.

From the information provided in the question, $P(\text{Spanish}) = 0.4$. Previously, we calculated $P(\text{Spanish}|\text{German})$:

$$\begin{aligned}
 P(\text{Spanish}|\text{German}) &= \frac{P(\text{Spanish and German})}{P(\text{German})} \\
 &= \frac{0.08}{0.2} \\
 &= 0.4
 \end{aligned}$$

We can see that $P(\text{Spanish}) = P(\text{Spanish}|\text{German})$. Because these two probabilities are equal, the events "Spanish" and "German" are independent. This means that the probability a student is taking Spanish does not affect the probability a student is taking German.

EXAMPLE

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

Cell Phone Use	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

Are the events "cell phone user" and "speeding violation in the last year" independent? Explain.

Solution

To check for independence, we need to check two probabilities: $P(\text{cell phone})$ and $P(\text{cell phone}|\text{speeding violation})$. If these probabilities are equal, the events are independent. If the probabilities are not equal, the events are dependent.

$$\begin{aligned} P(\text{cell phone}) &= \frac{305}{755} \\ &= 0.4040 \end{aligned}$$

$$\begin{aligned} P(\text{cell phone}|\text{speeding violation}) &= \frac{25}{70} \\ &= 0.03571 \end{aligned}$$

We can see that $P(\text{cell phone}) \neq P(\text{cell phone}|\text{speeding violation})$. Because these probabilities are not equal, the events "cell phone user" and "speeding violation" are dependent. This means that the probability a person is a cell phone user does affect the probability the person had a speeding violation in the last year.

TRY IT

At a local basketball game, 70\% of the fans are cheering for the home team, 25\% of the fans are wearing blue, and 12\% of the fans are cheering for the home team and wearing blue. Are the events "cheering for the home team" and "wearing blue" independent? Explain.

Click to see Solution

Because $P(\text{home team}) = 0.7$ does not equal $P(\text{home team}|\text{blue}) = \frac{0.12}{0.25} = 0.48$, the events "cheering for the home team" and "wearing blue" are dependent.

TRY IT

This table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

Stretching Practice	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

Are the events "does not stretch" and "injury in last year" independent? Explain.

Click to see Solution

Because $P(\text{no stretch}) = \frac{450}{800} = 0.5625$ does not equal $P(\text{no stretch}|\text{injury}) = \frac{231}{286} = 0.8077$, the events "does not stretch" and "injury in last year" are dependent.

Sampling may be done **with replacement** or **without replacement**, which effects whether or not events are considered independent or dependent.

- **With replacement:** If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent because the result of the first pick will not change the probabilities for the second pick.
- **Without replacement:** When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. Depending on the situation, the events are considered to be dependent or not independent.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=94#oembed-1>

Video: "Calculating conditional probability | Probability and Statistics | Khan Academy" by Khan Academy [6:43] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=94#oembed-2>

Video: "Conditional probability and independence | Probability | AP Statistics | Khan Academy" by Khan Academy [4:07] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A recent survey asked people about home ownership and annual income. A total of 750 people were surveyed. Of the 750 people surveyed, 425 owned a home, 338 people had an annual income of \$60,000 or more, and 293 people owned a home and had an annual income of \$60,000 or more.
 - a. what is the probability that one of the people in the survey owned a home, given that they had an annual income of \$60,000 or more?
 - b. Are the events "owned a home" and "annual income of \$60,000 or more" independent? Explain.

Click to see Answer

- a. 0.8669
- b. Dependent because

$$P(\text{owned home}) \neq P(\text{owned home} | \text{annual income of \$60,000 or more}).$$

2. A local college surveyed its recent graduates about their overall satisfaction with their college experience and employment status post-graduation. In the survey, 75\% of respondents said they were satisfied with their college experience, 64\% of respondents said they found full-time jobs after graduation, and 52\% of respondents said they were satisfied with their college experience and found full-time jobs after graduation.
- What is the probability that a respondent was satisfied with their college experience, given that they found full-time jobs after graduation?
 - Are the events "satisfied with college experience" and "found full-time job" independent? Explain.

Click to see Answer

- 0.8125
- Dependent because

$$P(\text{cell satisfied with college experience}) \neq P(\text{satisfied with college experience} | \text{found full-time job})$$

3. U and V are mutually exclusive events. $P(U) = 0.26$; $P(V) = 0.37$. Find $P(U|V)$

Click to see Answer

0

4. At a local college, 20\% of the students are studying business, 40\% of the students are studying mathematics, and 8\% of the students are studying both business and mathematics.
- What is the probability that a randomly selected student studies business, given that they study mathematics?
 - What is the probability that a randomly selected business student studies mathematics?
 - Are the events "business" and "mathematics" independent? Explain.

Click to see Answer

- 0.2
- 0.4
- Independent because $P(\text{business}) = P(\text{business} | \text{mathematics})$.

5. In a collection of eight cards, five cards are green, and three cards are yellow. The five green

cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. One card is selected at random.

- What is the probability the card is green, given that the card has an even number on it?
- Are the events "green" and "even" independent? Explain.
- What is the probability that a yellow card has an odd number on it?
- Are the events "yellow" and "odd" independent? Explain.

Click to see Answer

- 0.4
- Dependent because $P(\text{green}) \neq P(\text{green}|\text{even})$.
- 0.6667
- Dependent because $P(\text{odd}) \neq P(\text{odd}|\text{yellow})$.

- At a college, 72\% of courses have final exams, 46\% of courses require research papers, and 32\% of courses have both a research paper and a final exam.
 - Find the probability that a course has a final exam, given that it has a research project.
 - Are the events "final exam" and "research project" independent? Explain.

Click to see Answer

- 0.6957
- Dependent because $P(\text{final exam}) \neq P(\text{final exam}|\text{research project})$.

- In a box of assorted cookies, 36\% contain chocolate, 12\% contain nuts, and 8\% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.
 - If Sean selects a cookie that contains chocolate, what is the probability that the cookie contains nuts?
 - Are the events "chocolate" and "nuts" independent? Explain.

Click to see Answer

- 0.2222
- Dependent because $P(\text{nuts}) \neq P(\text{nuts}|\text{chocolate})$.

- A previous year, the weights, in pounds, of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the SAN JOSE MERCURY NEWS. The factual data are compiled into the table.

Shirt Number	At most 210	211–250	251–290	More than 290	Total
1–33	21	5	0	0	26
34–66	6	18	7	4	35
67–99	6	12	22	5	45
Total	33	35	29	9	106

For the following, suppose that one player from these two teams is selected at random.

- What is the probability that the player's weighs at most 210 pounds, given that the player has a shirt number in the 67-99 category?
- What is the probability that the player in the 34-66 category weighs between 251 and 290 pounds?
- If a player weighs more than 290 pounds, what is the probability the player has a shirt number in the 34-66 category?
- Are the events "1-33" and "211-250" independent? Explain

Click to see Answer

- 0.1333
- 0.2
- 0.4444
- Dependent because $P(1-33) \neq P(1-33|211-250)$.

- The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home Run	Total Hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

- What is the probability that a hit was made by Jackie Robinson, given that the hit was a double?
- If the hit was made by Hank Aaron, what is the probability that the hit was a single?
- What is the probability that a triple was hit by Babe Ruth?

- d. Are the events "Ty Cobb" and "home run" independent? Explain.

Click to see Answer

- a. 0.1731
- b. 0.6083
- c. 0.2333
- d. Dependent because $P(\text{Ty Cobb}) \neq P(\text{Ty Cobb}|\text{home run})$.

10. The table shows a random sample of musicians and how they learned to play their instruments.

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

- a. Find the probability a musician is male, given that they are self-taught.
- b. Find the probability that a female musician studied in school.
- c. If the musician had private lessons, find the probability the musician is female.
- d. Are the events "male" and "studied in school" independent? Explain.

Click to see Answer

- a. 0.6129
- b. 0.5278
- c. 0.5946
- d. Dependent because $P(\text{male}) \neq P(\text{male}|\text{studied in school})$.

11. The table below identifies a group of children by one of four hair colours and by type of hair.

Hair Type	Brown	Blond	Black	Red	Totals
Wavy	20	5	15	3	43
Straight	80	15	65	12	172
Totals	100	20	80	15	215

- a. What is the probability that a randomly selected child has wavy hair, given that they have black hair?

- b. What is the probability that a randomly selected child with straight hair has blond hair?
- c. If a child has red hair, what is the probability the child has wavy hair?
- d. Are the events "straight" or "brown" independent? Explain.

Click to see Answer

- a. 0.1875
- b. 0.0872
- c. 0.2
- d. Independent because $P(\text{straight}) = P(\text{straight}|\text{brown})$.

12. E and F are mutually exclusive events. $P(E) = 0.4$ and $P(F) = 0.5$. Find $P(E | F)$.

Click to see Answer

0

13. J and K are independent events. $P(J|K) = 0.3$. Find $P(J)$.

Click to see Answer

0.3

14. Q and R are independent events. $P(Q) = 0.4$ and $P(Q \text{ and } R) = 0.1$. Find $P(R)$.

Click to see Answer

0.25

15. Suppose $P(C) = 0.4$, $P(D) = 0.5$ and $P(C|D) = 0.6$.

- a. Find $P(C \text{ and } D)$.
- b. Are C and D mutually exclusive? Why or why not?
- c. Are C and D independent events? Why or why not?
- d. Find $P(C \text{ or } D)$.
- e. Find $P(D|C)$.

Click to see Answer

- a. 0.3
- b. No because $P(C \text{ and } D) \neq 0$
- c. Dependent because $P(C) \neq P(C|D)$.
- d. 0.6

e. 0.75

16. A college finds that 10\% of students have taken a distance learning class and that 40\% of students are part-time students. Of the part-time students, 20\% have taken a distance learning class.
- Find the probability that a student takes a distance learning class and is a part-time student.
 - Find the probability that a student is a part-time student, given that they take a distance learning class.
 - Find the probability that the student is a part-time student or takes a distance learning class.
 - Are the events "distance learning" and "part-time" independent? Explain.

Click to see Answer

- 0.08
- 0.8
- 0.42
- Dependent because $P(\text{distance learning}) \neq P(\text{distance learning}|\text{part-time})$.

"3.6 Conditional Probability" and "3.8 Exercises" from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

3.6 JOINT PROBABILITIES

LEARNING OBJECTIVES

- Calculate joint probabilities.

A **joint probability** is the probability of events A and B happening at the same time. We are interested in both events occurring simultaneously in the unrestricted sample space. We have seen these types of probabilities already when we looked at contingency tables and in the context of “or” probabilities.

EXAMPLE

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

Cell Phone Use	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

1. What is the probability that a randomly selected person is a cell phone user and had no speeding violations in the last year?
2. What is the probability that a randomly selected person had a speeding violation in the last year and does not use a cell phone?

Solution

$$1. \text{ Probability} = \frac{\text{number of cell phone users and no violations}}{\text{total number in study}} = \frac{280}{755}$$

$$2. \text{ Probability} = \frac{\text{number of violations and not cell phone users}}{\text{total number in study}} = \frac{45}{755}$$

NOTE

These two probabilities are examples of joint probabilities. For example, in part 1, we want to find the probability that a randomly selected person has both traits at the same time: cell phone user and no speeding violations. So, we are interested in both events happening simultaneously.

Repeated Trial Experiments

So far, most of the probabilities we have looked at are based on a **single trial experiment** and finding a probability based on that single trial. For example, finding the probability of rolling an even number in a single roll of a die is a single trial experiment—we are only rolling the die one time, and then we want to find the probability of a particular event happening in that single roll. Even the joint probabilities that we have seen so far, as in the example above, are based on a single trial experiment. We see these types of joint probabilities when we randomly select a single item and then want to find the probability that the item has two different characteristics at the same time.

However, we often want to calculate probabilities associated with **repeated trial experiments**. In a repeated trial experiment, we deal with **identical trials** that are repeated a number of times.

For example, flipping a coin three times is an example of a repeated trial experiment—the trial is flipping the coin, and then that trial is repeated three identical times.

EXAMPLE

Which of the following are repeated trial experiments? For the repeated trial experiments, identify the trial and the number of repetitions.

1. Finding the probability of rolling an odd number in the roll of the die.
2. Finding the probability of drawing five spades from a deck of cards.
3. Finding the probability a randomly selected person has blue eyes and blond hair.
4. Finding the probability that three women from a pool of candidates are selected for a committee.
5. Finding the probability that a student answers ten multiple choice questions correctly.

Solution

1. Single trial experiment. The die is rolled one time.
2. Repeated trial experiment. The trial is selecting a card from the deck and this trial is repeated five times.
3. Single trial experiment. A single person is selected.
4. Repeated trial experiment. The trial is selecting a woman from the candidate pool, and this trial is repeated three times.
5. Repeated trial experiment. The trial is answering an individual question, and this trial is repeated ten times.

We can think of repeated trial experiments as joint probabilities—event on trial one AND event on trial two AND event on trial three, and so on, depending on the number of trials. Suppose in the example of flipping the coin three times, we want to find the probability of getting three heads in the three flips. We can think of this as a joint probability—heads on flip one AND heads on flip two AND heads on flip three. We want to calculate probabilities for such repeated trial experiments and, as we will see, the key to such probabilities is to think of the repeated trials as a joint probability.

One thing we must consider in a repeated trial experiment is whether the trials are done with or

without replacement because this changes how we calculate the probability as we move from trial to trial.

- **With replacement.** Each member of a population is replaced after it is selected on a trial, and so each member of the population has the possibility of being chosen more than once (on different trials of the experiment). In terms of probability, with replacement means that the probability a member of the population is chosen stays the same from trial to trial. In other words, the trials are independent events because the result of the first trial does not affect the result of the second trial.
- **Without replacement.** Each time a member of a population is selected, it is NOT replaced, and so each member of the population cannot be chosen more than once. In terms of probability, without replacement means that the probability a member of the population is chosen changes from trial to trial. In other words, the trials are dependent events because the result of the first trial does affect the result of the second trial.

When calculating probabilities for repeated trial experiments, it is important that we identify if the experiment is done with or without replacement. Sometimes, we will be told directly that the experiment is done with or without replacement. But most of the time we will need to determine if the experiment is done with or without replacement from the context of the question.

EXAMPLE

For each of the following, determine if the experiment is done with or without replacement.

1. Flipping a coin three times.
2. Selecting three women for a committee from a pool of candidates.
3. Drawing five cards from the deck of cards.
4. Rolling a die six times.
5. Selecting the members of the student executive committee from the student council.

Solution

1. With replacement. The probability of heads or tails stays the same with each flip.

2. Without replacement. In this case, we want three different women on the committee, so we must select them without replacement. (Selecting with replacement would mean a possibility of the same woman being selected three times, and then the committee would consist of just a single person).
3. Depending on the context, this could be with or without replacement. If each card is replaced after it is selected, this would be with replacement. If each card is not replaced after it is selected, this would be without replacement. In this situation, the question would probably include a statement about whether the cards are drawn with or without replacement.
4. With replacement. The probability of rolling any of the numbers stays the same with each roll of the die.
5. Without replacement. In this case, we want the members of the executive committee to be all different, so we must select them without replacement.

The Multiplication Rule for Joint Probabilities

In mathematical terms, “and” means multiply. By thinking of a repeated trial experiment as a joint probability, the basic idea is to multiply the probabilities of the individual trials. Basically, if we think of a repeated trial experiment as a joint probability—event on trial one AND event on trial two AND event on trial three and so on, depending on the number of trials—we can find the probability by multiplying together the probabilities of the trials:

$$\text{Probability} = \text{Prob. on Trial 1} \times \text{Prob. on Trial 2} \times \text{Prob. on Trial 3} \times \cdots$$

Unfortunately, it is more complicated than that because we have to work out the probabilities on each trial, and these probabilities are affected by whether the experiment is done with or without replacement.

The multiplication rule to find the probability of A and B in a repeated trial experiment is

$$P(A \text{ and } B) = P(A) \times P(B|A)$$

If we think of A as the first trial and B as the second trial, the probability of A and B is the probability of A (the probability of A on the first trial) times the probability of B given A (the probability of B on the second trial **assuming** that A happened on the first trial).

In the case that the experiment is done with replacement, the events A and B are **independent**, so $P(B|A) = P(B)$ and this rule becomes

$$P(A \text{ and } B) = P(A) \times P(B)$$

We can extend this rule to any number of trials, we just need to keep multiplying as we move from trial to trial.

When finding probabilities associated with repeated trial experiments, remember the following:

- To find the probability, we work with the probabilities of the individual trials, multiplying the probabilities together as we move from trial to trial.
- Identify if the experiment is done with or without replacement and use that information to find the probability on each subsequent trial.

EXAMPLE

A small local high school has 25 students in its graduating class. 18 of the students are going to college next year, and the remaining 7 are not going to college next year. Suppose two students are selected at random from the graduating class.

1. What is the probability that both students are going to college next year?
2. What is the probability that exactly one of the students is going to college next year?

Solution

This is a repeated trial experiment. A trial is selecting a student, and there are two trials. The assumption here is that the experiment is done without replacement because we do not want to get the same student twice.

1. We want to get college-bound students on both trials. In other words, college-bound student on trial one AND college-bound student on trial two. On the first trial, the probability of getting a college-bound student is $\frac{18}{25}$. We are selecting without replacement, so after the first trial, we assume that we have removed one of the college-bound students. This means

that on the second trial, there are only 24 students to pick from (one student was removed on the first trial,) and there are only 17 college-bound students left (on the first trial we removed one of the 18 college-bound students). So on the second trial, the probability of getting a college-bound student is $\frac{17}{24}$. The probability of getting two college-bound students is

$$\text{Probability} = \frac{18}{25} \times \frac{17}{24} = 0.51$$

2. We want one college-bound student (denoted C) and one non-college-bound student (denoted N). In this case, we have to think about the **order** of the selections—there is a difference between college-bound on trial one, non-college-bound on trial two (CN) **and** non-college-bound on trial one, college-bound on trial two (NC). All possible orders must be accounted for when we calculate the probability. One of the two possible orders must occur: CN OR NC . For each of the individual orders, we multiply the probabilities as we move from trial to trial. The “or” means that we add the probabilities of the different orders. In other words:

$$\text{Probability} = \text{Probability of } CN + \text{Probability of } NC$$

For the CN order (college-bound on trial one, non-college-bound on trial two), we want a college-bound student on trial one, and the probability of getting a college-bound student is $\frac{18}{25}$. We are selecting without replacement, so after the first trial, we assume that we have removed one of the college-bound students. This means that on the second trial, there are only 24 students to pick from (a college-bound student was removed on the first trial), and there are 7 non-college-bound students (none of the non-college-bound students were removed after the first trial). So on the second trial, the probability of getting a non-college-bound student is $\frac{7}{24}$. So the probability of getting the CN order is $\frac{18}{25} \times \frac{7}{24}$.

Similarly, for the NC order (non-college-bound on trial one, college-bound on trial two), we want a non-college-bound student on trial one, and the probability of getting a non-college-bound student is $\frac{7}{25}$. We are selecting without replacement, so after the first trial, we assume that we have removed one of the non-college-bound students. This means that on the second trial, there are only 24 students to pick from (a non-college-bound student was removed on the first trial), and there are 18 college-bound students (none of the college-bound students

were removed after the first trial). So on the second trial, the probability of getting a college-bound student is $\frac{18}{24}$. So the probability of getting the NC order is $\frac{7}{25} \times \frac{18}{24}$.

The probability of getting exactly one college-bound student is

$$\begin{aligned}\text{Probability} &= \text{Probability of } CN + \text{Probability of } NC \\ &= \left(\frac{18}{25} \times \frac{7}{24} \right) + \left(\frac{7}{24} \times \frac{18}{25} \right) \\ &= 0.42\end{aligned}$$

EXAMPLE

Suppose a fair die is rolled two times.

1. What is the probability of getting two 5's?
2. What is the probability of getting exactly one 2 and one 6 in the two rolls?

Solution

This is a repeated trial experiment. A trial is rolling a die, and there are two trials. The trials are independent (what happens on the first roll does not affect what happens on the second roll).

1. We want to get a 5 on both rolls. In other words, a 5 on roll one AND a 5 on roll two. On the first roll, the probability of getting a 5 is $\frac{1}{6}$. On the second roll, the probability of getting a 5 is $\frac{1}{6}$. Because the rolls are independent, the probability of getting a 5 on the second roll is not affected by what happens on the first roll. The probability of getting two 5's is

$$\text{Probability} = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

2. We want one 2 and one 6. In this case, we have to think about the **order** of the rolls—there is a

difference between 2 on roll one, 6 on roll two **and** 6 on roll one, 2 on roll two. All possible orders must be accounted for when we calculate the probability. One of the two possible orders must occur: 26 OR 62. For the individual orders, we multiply the probabilities as we move from trial to trial. The “or” means that we add the probabilities of the different orders. In other words,

$$\text{Probability} = \text{Probability of 26} + \text{Probability of 62}$$

For the 26 order (2 on roll one, 6 on roll two), the probability of getting a 2 on roll one is $\frac{1}{6}$ and the probability of getting a 6 on roll two is $\frac{1}{6}$. So the probability of getting the 26 order is $\frac{1}{6} \times \frac{1}{6}$.

Similarly for the 62 order (6 on roll one, 2 on roll two), the probability of getting a 6 on roll one is $\frac{1}{6}$ and the probability of getting a 2 on roll two is $\frac{1}{6}$. So the probability of getting the 62 order is $\frac{1}{6} \times \frac{1}{6}$.

The probability of getting exactly one 2 and one 6 is

$$\begin{aligned} \text{Probability} &= \text{Probability of 26} + \text{Probability of 62} \\ &= \left(\frac{1}{6} \times \frac{1}{6} \right) + \left(\frac{1}{6} \times \frac{1}{6} \right) \\ &= \frac{1}{18} \end{aligned}$$

TRY IT

A box contains 30 microchips, and 9 of those microchips are defective. Suppose two microchips are randomly selected from the box for inspection by the quality control officer.

1. What is the probability that both microchips are defective?

2. What is the probability that exactly one of the microchips is defective?

Click to see Solution

1. $\text{Probability} = \frac{9}{30} \times \frac{8}{29} = 0.0828$
2. $\text{Probability} = \left(\frac{9}{30} \times \frac{21}{29} \right) + \left(\frac{21}{30} \times \frac{9}{29} \right) = 0.4345$

EXAMPLE

A box contains 5 red cards and 12 white cards. Suppose three cards are drawn at random from the box **without** replacement.

1. What is the probability that all three cards are white?
2. What is the probability that exactly one of the cards is red?
3. What is the probability that at least one card is red?
4. What is the probability that, at most, one card is white?

Solution

This is a repeated trial experiment. A trial is selecting a card, and there are three trials. There are 17 cards in the box.

1. We want to get a white card on all three draws. In other words, white on draw one AND white on draw two AND white on draw three. On the first draw, the probability of getting a white card is $\frac{12}{17}$. We are selecting without replacement, so after the first draw, we assume that we removed a white card from the box. This means that on the second draw, there are only 16 cards left in the box (one card was removed on the first draw), and there are only 11 white cards left (on the first draw we removed one of the 12 white cards). So on the second draw, the

probability of getting a white card is $\frac{11}{16}$. After the second draw we assume that we removed white cards from the box on draws one and two. This means that on the third draw, there are only 15 cards left in the box (two cards were removed on the first two draws), and there are only 10 white cards left (white cards were removed on draws one and two). So on the third draw, the probability of getting a white card is $\frac{10}{15}$. The probability of getting three white cards is

$$\text{Probability} = \frac{12}{17} \times \frac{11}{16} \times \frac{10}{15} = 0.3235$$

2. We want one red card (R), so the other two cards must be white (W). In this case, we have to think about the **order** of the selection. All possible orders must be accounted for when we calculate the probability. One of three possible orders must occur: RWW OR WRW OR WWR . For each of the individual orders, we multiply the probabilities as we move from draw to draw. The “or” means that we add the probabilities of the different orders. In other words,

$$\text{Probability} = P(RWW) + P(WRW) + P(WWR)$$

For the RWW order, the probability of red on draw one is $\frac{5}{17}$. We are selecting without replacement, so after the first trial, we assume that we have removed one of the red cards. This means that on the second draw, there are only 16 cards left in the box (one card was removed on the first draw) and all 12 white cards are left (on the first draw we removed a red card). So on the second draw, the probability of getting a white card is $\frac{12}{16}$. After the second draw, we assume that we removed a red card on draw one and a white card on draw two. This means that on the third draw, there are only 15 cards left in the box (two cards were removed on the first two draws), and there are only 11 white cards left (one white card was removed on draw two). So on the third draw, the probability of getting a white card is $\frac{11}{15}$. So the probability of getting the RWW order is $\frac{5}{17} \times \frac{12}{16} \times \frac{11}{15}$.

Using similar logic, the probability of getting the WRW order is $\frac{12}{17} \times \frac{5}{16} \times \frac{11}{15}$ and the probability of getting the WWR order is $\frac{12}{17} \times \frac{11}{16} \times \frac{5}{15}$.

The probability of getting exactly one red card is

$$\begin{aligned}\text{Probability} &= P(RWW) + P(WRW) + P(WWR) \\ &= \left(\frac{5}{17} \times \frac{12}{16} \times \frac{11}{15} \right) + \left(\frac{12}{17} \times \frac{5}{16} \times \frac{11}{15} \right) + \left(\frac{12}{17} \times \frac{11}{16} \times \frac{5}{15} \right) \\ &= 0.4853\end{aligned}$$

3. We want at least one red card in the three draws. This means we can have exactly one red card, or exactly two red cards, or exactly three red cards. As before, we have to think about the **order** of the selection. All possible orders must be accounted for when we calculate the probability. Here, there are seven possible ways of getting at least one red card: *RWW* OR *WRW* OR *WWR* OR *RRW* OR *RWR* OR *RRW* OR *RRR*. Of course, we could work out the probabilities of each of these orders and add them all up. But there is a faster way to find this probability—use the complement. The complement of “at least one red card” is “exactly zero red cards.” When we look at the seven possible orders that make up the “at least one red card” event, the complement consists of all of the missing orders. In this case, there is only one missing order, *WWW*, which is the event “exactly zero red cards.” Using the complement, the probability of at least one red card is

$$\begin{aligned}P(\text{at least one red card}) &= 1 - P(\text{exactly zero red card}) \\ &= 1 - P(WWW)\end{aligned}$$

In part 1 of this question, we found the probability of *WWW*: $\frac{12}{17} \times \frac{11}{16} \times \frac{10}{15}$. So, the probability of at least one red card is

$$\begin{aligned}P(\text{at least one red card}) &= 1 - P(WWW) \\ &= 1 - \left(\frac{12}{17} \times \frac{11}{16} \times \frac{10}{15} \right) \\ &= 0.6765\end{aligned}$$

4. We want at most one white card in the three draws. This means we can have exactly zero white cards or exactly one white card. As before, we have to think about the **order** of the selection. All possible orders must be accounted for when we calculate the probability. Here, there are four possible ways of getting at most one white card: *RRR* OR *RRW* OR *RWR* OR *WRR*. Using similar logic to above, the probability of getting the *RRR* order is $\frac{5}{17} \times \frac{4}{16} \times \frac{3}{15}$, the probability of getting the *RRW* order is $\frac{5}{17} \times \frac{4}{16} \times \frac{12}{15}$, the

probability of getting the RWR order is $\frac{5}{17} \times \frac{12}{16} \times \frac{5}{15}$, and the probability of getting the WRR order is $\frac{12}{17} \times \frac{5}{16} \times \frac{4}{15}$. The probability of getting at most one white card is

$$\begin{aligned} \text{Probability} &= P(RRR) + P(RRW) + P(RWR) + P(WRR) \\ &= \left(\frac{5}{17} \times \frac{4}{16} \times \frac{3}{15} \right) + \left(\frac{5}{17} \times \frac{4}{16} \times \frac{12}{15} \right) + \left(\frac{5}{17} \times \frac{12}{16} \times \frac{4}{15} \right) \\ &\quad + \left(\frac{12}{17} \times \frac{5}{16} \times \frac{4}{15} \right) \\ &= 0.1912 \end{aligned}$$

TRY IT

A box contains 5 red cards and 12 white cards. Suppose three cards are drawn at random from the box **with** replacements.

1. What is the probability that all three cards are white?
2. What is the probability that exactly one of the cards is red?
3. What is the probability that at least one card is red?
4. What is the probability that, at most, one card is white?

Click to see Solution

$$1. \text{ Probability} = \frac{12}{17} \times \frac{12}{17} \times \frac{12}{17} = 0.3517$$

$$2. \text{ Probability} = \left(\frac{5}{17} \times \frac{12}{17} \times \frac{12}{17} \right) + \left(\frac{5}{17} \times \frac{12}{17} \times \frac{12}{17} \right) + \left(\frac{5}{17} \times \frac{12}{17} \times \frac{12}{17} \right) = 0.4397$$

$$3. \text{ Probability} = 1 - \left(\frac{12}{17} \times \frac{12}{17} \times \frac{12}{17} \right) = 0.6483$$

$$4. \text{ Probability} = \left(\frac{5}{17} \times \frac{5}{17} \times \frac{5}{17} \right) + \left(\frac{5}{17} \times \frac{5}{17} \times \frac{12}{17} \right) + \left(\frac{5}{17} \times \frac{12}{17} \times \frac{5}{17} \right) + \left(\frac{12}{17} \times \frac{5}{17} \times \frac{5}{17} \right) = 0.2086$$

EXAMPLE

A company produces a popular brand of sports drink. The company is currently running a contest where winning symbols are placed under the bottle caps. 7% of all the bottle caps contain winning symbols. You buy three bottles of the sports drink.

1. What is the probability that all bottles have winning symbols?
2. What is the probability that exactly one of the bottles has a winning symbol?
3. What is the probability that at least one bottle has a winning symbol?

Solution

This is a repeated trial experiment. A trial is selecting a bottle, and there are three trials. This is an experiment without replacement (you do not want to select the same bottle three times). However, because the population of bottles is very, very large, we can treat the experiment as if the selections are made with replacements. This means that we can treat the selection of the bottles as independent, and so the probability of getting a winning bottle will be 7% on every draw.

1. We want to get a winning symbol on all three bottles. In other words, win on bottle one AND win on bottle two AND win on bottle three. The probability of winning on the first bottle is 7%, the probability of winning on the second bottle is 7%, and the probability of winning on the third bottle is 7%. Because we can treat the selections as independent, the probability of winning does not change from draw to draw. The probability of getting three winning bottles is

$$\text{Probability} = 0.07 \times 0.07 \times 0.07 = 0.0003$$

2. We want one winning bottle (W), so the other two bottles must be non-winners (N). The probability of winning on any bottle is 7%, so the probability of losing on any bottle is 93%. In this case, we have to think about the **order** of the selection. All possible orders must be accounted for when we calculate the probability. One of the three possible orders must occur: WNN OR NWN OR NNW . For each of the individual orders, we multiply the probabilities as we move from draw to draw. The “or” means that we add the probabilities of the different orders. In other words,

$$\text{Probability} = P(WNN) + P(NWN) + P(NNW)$$

For the WNN order (win on bottle one, non-wins on bottles two and three), the probability of getting a win on bottle one is 7%, and the probability of getting a non-win on bottle two or bottle three is 93%. So the probability of getting the WNN order is $0.07 \times 0.93 \times 0.93$. Using similar logic, the probability of getting the NWN order is $0.93 \times 0.07 \times 0.93$ and the probability of getting the NNW order is $0.93 \times 0.93 \times 0.07$.

The probability of getting exactly one winning bottle is

$$\begin{aligned} \text{Probability} &= P(WNN) + P(NWN) + P(NNW) \\ &= (0.07 \times 0.93 \times 0.93) + (0.93 \times 0.07 \times 0.93) + (0.93 \times 0.93 \times 0.07) \\ &= 0.1816 \end{aligned}$$

3. We want at least one winning bottle. This means we can have exactly one winning bottle, or exactly two winning bottles, or exactly three winning bottles. Of course, we could work out the probabilities of each of these orders and add them all up. But the faster way to find this probability is to use the complement. The complement of “at least one winning bottle” is “exactly zero winning bottles.” The “exactly zero winning bottle” is the case NNN (all three bottles are non-winners). Using the complement, the probability of at least one winning bottle is

$$\begin{aligned} P(\text{at least one winner}) &= 1 - P(\text{exactly zero winners}) \\ &= 1 - P(NNN) \end{aligned}$$

The probability of zero winning bottles (NNN) is: $0.93 \times 0.93 \times 0.93$. So, the probability of at least one winning bottle is

$$\begin{aligned}
 P(\text{at least one winner}) &= 1 - P(NNN) \\
 &= 1 - (0.93 \times 0.93 \times 0.93) \\
 &= 0.1956
 \end{aligned}$$

NOTE

In situations like the previous example, where we are drawing with replacement from a very, very large population, we treat the draws as if they are independent. Because the population is so large, the change in the probability as we go from draw to draw is very, very small, which makes it hardly detectable in the calculation of the answer. In such situations, we can treat the draws as independent. We cannot do this when we are drawing without replacement from a small population (as in the red and white card example above) because there are distinct changes in the probabilities as we move from draw to draw.

Exercises

1. A box contains 6 red marbles and 4 blue marbles. Suppose two marbles are drawn successively without replacement.
 - a. What is the probability that both marbles are red?
 - b. What is the probability that one marble is red and the other marble is blue?

Click to see Answer

- a. 0.3333
 - b. 0.5333
2. A box contains 13 red marbles and 8 blue marbles. Suppose two marbles are drawn successively with replacement.
 - a. What is the probability that both marbles are blue?

- b. What is the probability that one marble is red and the other marble is blue?

Click to see Answer

- a. 0.1451
- b. 0.4717

3. Based on the results of a survey, 60\% of the population likes chocolate ice cream. Suppose three people are selected at random.

- a. What is the probability that all three people like chocolate ice cream?
- b. What is the probability that exactly one person likes chocolate ice cream?
- c. What is the probability that at least one person likes chocolate ice cream?
- d. What is the probability that at most one person likes chocolate ice cream?

Click to see Answer

- a. 0.216
- b. 0.288
- c. 0.936
- d. 0.352

4. A box of cookies contains three chocolate and seven butter cookies. Miguel randomly selects two cookies from the box.

- a. What is the probability that both cookies are chocolate cookies?
- b. What is the probability that both cookies are the same flavour?
- c. What is the probability that the cookies are different flavours?

Click to see Answer

- a. 0.0667
- b. 0.5333
- c. 0.4667

5. A company has 25 employees. Seven of the employees are management, and the rest of the employees are in non-management positions. Three employees are selected at random to be on the safety oversight committee.

- a. What is the probability that all three employees are non-management?
- b. What is the probability that exactly one of the employees is management?
- c. What is the probability that at least one of the employees is non-management?
- d. What is the probability that at most one of the employees is management?

Click to see Answer

- a. 0.3548
- b. 0.4657
- c. 0.9848
- d. 0.8204

6. Data shows that 43\% of the population has blood type O.

- a. Suppose three people are selected at random. What is the probability that all three people have blood type O?
- b. Suppose five people are selected at random. What is the probability that all five people do not have blood type O?
- c. Suppose four people are selected at random. What is the probability that exactly one person as blood type O?
- d. Suppose three people are selected at random. What is the probability that exactly two people have blood type O?
- e. Suppose four people are selected at random. What is the probability that at least one person has blood type O?
- f. Suppose three people are selected at random. What is the probability that at most one person does not have blood type O?

Click to see Answer

- a. 0.0795
- b. 0.0602
- c. 0.3185
- d. 0.3162
- e. 0.8944
- f. 0.3957

7. A box contains 40 smartphones, and five of those smartphones are defective.

- a. Suppose three smartphones are selected at random. What is the probability that all three smartphones are defective?
- b. Suppose five smartphones are selected at random. What is the probability that all five smartphones are not defective?
- c. Suppose four smartphones are selected at random. What is the probability that exactly one of the smartphones is defective?
- d. Suppose five smartphones are selected at random. What is the probability that exactly one of the smartphones is not defective?
- e. Suppose three smartphones are selected at random. What is the probability that at most

one of the smartphones is defective?

- f. Suppose four smartphones are selected at random. What is the probability that at least one of the smartphones is defective?

Click to see Answer

- a. 0.001
- b. 0.4934
- c. 0.3581
- d. 0.0003
- e. 0.9636
- f. 0.4271

8. A company produces microchips. Based on previous analysis, the company knows that 4% of the chips are defective.

- a. Suppose four chips are randomly selected. What is the probability that all four chips are not defective?
- b. Suppose three chips are randomly selected. What is the probability that exactly one chip is defective?
- c. Suppose four chips are randomly selected. What is the probability that exactly three chips are not defective?
- d. Suppose three chips are randomly selected. What is the probability that at least one chip is defective?
- e. Suppose four chips are randomly selected. What is the probability that, at most, one chip is not defective?

Click to see Answer

- a. 0.8493
- b. 0.1106
- c. 0.1416
- d. 0.1153
- e. 0.0002

9. A 12-sided die is in the shape of a regular dodecahedron. The faces of the 12-sided die are labelled with the numbers 1 to 12. Suppose the 12-sided die is rolled three times.

- a. What is the probability that all three rolls are the number 7?
- b. What is the probability that all three rolls are even numbers?
- c. What is the probability that exactly one of the rolls is a multiple of 3?
- d. What is the probability that exactly two of the rolls are less than 6?

- e. What is the probability that at least one of the rolls is greater than 9?

Click to see Answer

- a. 0.0006
- b. 0.125
- c. 0.4444
- d. 0.3038
- e. 0.5781

10. A special deck of cards has fifteen cards. Eight are green, four are blue, and three are red.
- a. Suppose three cards are picked without replacement. What is the probability that all three cards are green?
 - b. Suppose three cards are picked without replacement. What is the probability that exactly two of the cards are blue?
 - c. Suppose three cards are picked without replacement. What is the probability that at least one of the cards is red?
 - d. Suppose three cards are picked without replacement. What is the probability that, at most, one of the cards is green?
 - e. Suppose three cards are picked with replacement. What is the probability that all three cards are green?
 - f. Suppose three cards are picked with replacement. What is the probability that exactly two of the cards are blue?
 - g. Suppose three cards are picked with replacement. What is the probability that at least one of the cards is red?
 - h. Suppose three cards are picked with replacement. What is the probability that, at most, one card is green?

Click to see Answer

- a. 0.1231
- b. 0.1451
- c. 0.5165
- d. 0.2
- e. 0.1517
- f. 0.1564
- g. 0.488
- h. 0.2178

11. A cup contains three red, four yellow and five blue beads.

- a. Suppose three beads are selected at random without replacement. What is the probability all three beads are blue?
- b. Suppose three beads are selected at random with replacement. What is the probability all three beads are blue?
- c. Suppose three beads are selected at random without replacement. What is the probability that exactly one of the beads is red?
- d. Suppose three beads are selected at random with replacement. What is the probability that exactly one of the beads is red?
- e. Suppose three beads are selected at random without replacement. What is the probability that at least one bead is yellow?
- f. Suppose three beads are selected at random with replacement. What is the probability that at least one bead is yellow?

Click to see Answer

- a. 0.0455
- b. 0.0723
- c. 0.4909
- d. 0.4219
- e. 0.7455
- f. 0.7037

“3.7 Joint Probabilities” and “3.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

PART IV

DISCRETE PROBABILITY DISTRIBUTIONS

Previously, we looked at probability in terms of a probability experiment, the outcomes of the experiment, and the associated probabilities of the outcomes of those experiments. These ideas can be extended to the idea of random variables and probability distributions. Random variables and their associated probability distributions allow us to study populations of data and answer probability questions about those populations.

For example, suppose a student takes a ten-question, true-false quiz. Because the student had such a busy schedule, they could not study and guesses randomly at each answer. What is the probability of the student passing the test with at least 70\% (7 out of 10 correct answer)? In this case, the random variable we are interested in is the number of correct answers the student got on the test, and we want to find the probability the student got 7 or more correct answers on the quiz.

As another example, suppose a company is interested in the number of long-distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long-distance phone calls during the peak time? Here, the random variable we are interested in is the number of long-distance phone calls made during the peak time and we want to find the probability the number of long-distance phone calls is greater than 20.

As seen in the above examples, a **random variable** describes the outcomes of a statistical experiment in words. The values a random variable takes on are numbers associated with the outcomes of the experiment. These two examples illustrate two different types of probability problems involving **discrete random variables**. Recall that discrete data are data that can be counted. In this chapter, we want to define random variables, construct the probability distribution for an associated random variable, and use the probability distribution to calculate probabilities for the random variable.

CHAPTER OUTLINE

4.1 Random Variables

4.2 Probability Distribution of a Discrete Random Variable

4.3 Expected Value and Standard Deviation for a Discrete Probability Distribution

4.4 The Binomial Distribution

4.5 The Poisson Distribution

“4.1 Introduction to Discrete Random Variables” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

4.1 RANDOM VARIABLES

LEARNING OBJECTIVES

- Describe a random variable.
- Identify discrete or continuous random variables.

Previously, we learned about probability experiments and the possible outcomes of the experiment (the sample space). For example, flipping a coin two times is an experiment, and the possible outcomes of that experiment are $\{HH, HT, TH, TT\}$. We also learned how to calculate probabilities associated with the outcomes of the experiment. In the case of flipping a coin two times, the probability of getting two heads is 25\%.

Instead of listing the outcomes of the experiment, we want to associate numerical values with the outcomes of an experiment. For the flipping the coin two times experiment, we can associate the number of heads we get on the two flips with the outcomes in the sample space: two heads is associated with the outcome HH , one head is associated with the outcomes HT and TH , and zero heads is associated with the outcome TT . In this way, we can assign a numerical value to the outcomes of the experiment. This numerical description of the outcomes of an experiment is called a **random variable**.

A **random variable** is a numerical description of the outcomes of an experiment. Each possible outcome of an experiment is associated with a numerical value based on the random variable. The numerical values the random variable takes on depends on how the variable is defined and the outcomes of the experiment. The values of a random variable can vary with each repetition of an experiment. Upper case letters such as X or Y denote a random variable. Lowercase letters like x or y denote the value of a random variable. If X is a random variable, then X is written in words, and x is given as a number. For example, if X is the number of children in a family, then x represents a specific integer $0, 1, 2, 3, \dots$.

EXAMPLE

Suppose a coin is tossed three times. Define a random variable for this experiment. List all the possible values of the random variable.

Solution

Define the random variable X as the number of tails when a coin is tossed three times.

The sample space for this experiment is

$\{TTT, THH, HTH, HHT, HTT, THT, TTH, HHH\}$. Then the possible values of the random variable are $x = 0, 1, 2, 3$. Here, $x = 0$ corresponds to getting exactly zero tails on the three flips (i.e. the outcome HHH), $x = 1$ corresponds to getting exactly one tail on the three flips (i.e. the outcomes THH, HTH, HHT), $x = 2$ corresponds to getting exactly two tails on the three flips (i.e. the outcomes TTH, THT, HTT), and $x = 3$ corresponds to getting exactly three tails on the three flips (i.e. the outcome TTT).

NOTES

1. The random variable X is described in words, and the values of x are numbers.
2. All of the outcomes in the sample space are associated with a value of x .
3. Instead of the number of tails in the three flips, another way to define a random variable for this experiment is the number of heads in the three flips.

TRY IT

Suppose a six-sided die is rolled a single time. Define a random variable for this experiment. List all possible values of the random variable.

Click to see Solution

Define the random variable X as the number on the top face of the die. The possible values of the random variable are $x = 1, 2, 3, 4, 5, 6$.

Variables in statistics differ from variables in intermediate algebra in the two following ways:

- The values a random variable can take on are not necessarily numbers. The values may be expressed in words. For example, if X is hair colour, then the values of the random variable are {black, blond, grey, brown, red}.
- We can tell what specific value, x , the random variable X takes only after performing the experiment.

Discrete Random Variables

A random variable that only takes on certain numerical values is called a **discrete random variable**. For example, the random variable defined as the number of heads obtained in two flips of a coin is a discrete random variable because the random variable can only take on the values 0, 1, and 2.

EXAMPLE

Consider the experiment of orders taken at a restaurant drive-thru window in a single day. Define a random variable for this experiment. List all possible values of the random variable.

Solution

Define the random variable X as the number of orders taken at the drive-thru window during a one-day period. The possible values of the random variable are $x = 0, 1, 2, 3, 4, \dots$

NOTES

This is a discrete random variable because the random variable can only take on a value from the numbers $0, 1, 2, 3, \dots$. The random variable is a count of the number of orders, and so must be a non-negative whole number. For example, the random variable cannot take on the value of **20.73** because it is not possible to take **20.73** orders at the drive-thru window.

Continuous Random Variables

A random variable that takes on any numerical value in an interval or a collection of intervals is called a **continuous random variable**. Examples of continuous random variables are random variables associated with measurements, such as time, weight, height, or distance. For example, the random variable defined as the height of students in a class is a continuous random variable because the height of a student may take on any positive number.

EXAMPLE

Consider the experiment of measuring the mass of a package shipped by the post office. Define a random variable for this experiment. List all possible values of the random variable.

Solution

Define the random variable X as the mass of a package. The random variable may take on any non-negative number.

NOTES

This is a continuous random variable because the random variable can take on any number greater than or equal to 0.

NOTE

The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable because you count the miles. If X is the distance you drive to work, then X is a continuous random variable because you measure the miles. For a second example, if X is equal to the number of books in a backpack, then X is a discrete random variable because the number of books is a count. If X is the weight of a book, then X is a continuous random variable because weights are measured. How the random variable is defined is very important.

TRY IT

Consider the experiment of preparing tax returns in an accounting office. Define a random variable for this experiment. List all possible values of the random variable. Identify the random variable as discrete or continuous.

Click to see Solution

Here are a couple of possible answers, depending on what we want the random variable to measure.

1. Define the random variable X as the time it takes to prepare a tax return. The random variable may take on any non-negative number. This is a continuous random variable.
2. Define the random variable Y as the number of tax returns the office can prepare in a day. The possible values of the random variable are $y = 0, 1, 2, 3, 4, \dots$. This is a discrete random variable.
3. Define the random variable Z as the amount of money owing on a tax return. The random variable may take on any possible dollar amount. This is a discrete random variable because the random variable cannot take on values such as 12.37563.

Exercises

1. A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer.
 - a. Define a random variable for this experiment.
 - b. List the possible values of the random variable.
 - c. Is the random variable discrete or continuous?

Click to see Answer

- a. The number of batches of muffins the baker needs to make.
- b. $0, 1, 2, 3, \dots$

c. Discrete.

2. A meteorologist monitors the temperature at the airport each day.

- a. Define a random variable for this experiment.
- b. List the possible values of the random variable.
- c. Is the random variable discrete or continuous?

Click to see Answer

- a. The temperature at the airport on any given day.
- b. Any possible number.
- c. Continuous.

3. Ellen is supposed to practice her music three days a week.

- a. Define a random variable for this experiment.
- b. List the possible values of the random variable.
- c. Is the random variable discrete or continuous?

Click to see Answer

- a. The number of times Ellen practices each week.
- b. 0, 1, 2, 3
- c. Discrete.

4. An IT helpdesk monitors the time a worker spends with a client on a call.

- a. Define a random variable for this experiment.
- b. List the possible values of the random variable.
- c. Is the random variable discrete or continuous?

Click to see Answer

- a. The amount of time, in minutes, the worker spends on a call.
- b. Any non-negative number.
- c. Continuous.

5. Javier volunteers in community events each month. He does not do more than five events in a month.

- a. Define a random variable for this experiment.
- b. List the possible values of the random variable.
- c. Is the random variable discrete or continuous?

Click to see Answer

- a. The number of events Javier does in a month.
 - b. 0, 1, 2, 3, 4, 5.
 - c. Discrete.
6. A quality control expert monitors the volume of paint in 4-litre paint cans on an assembly line. The quality control expert randomly selects a paint can off of the assembly line to check its volume
- a. Define a random variable for this experiment.
 - b. List the possible values of the random variable.
 - c. Is the random variable discrete or continuous?

Click to see Answer

- a. The number of litres of paint in the paint can.
- b. Any non-negative number.
- c. Continuous.

“4.1 Introduction to Discrete Random Variables“, “5.1 Introduction to Continuous Random Variables“, and “4.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

4.2 PROBABILITY DISTRIBUTION OF A DISCRETE RANDOM VARIABLE

LEARNING OBJECTIVES

- Recognize, understand, and construct discrete probability distributions.

The **probability distribution** for a random variable lists all the possible values of the random variable and the probability the random variable takes on each value. The probability distribution for a random variable describes how probabilities are distributed over the values of the random variable. A probability distribution can be a table, with a column for the values of the random variable and another column for the corresponding probability, or a graph, like a histogram with the values of the random variable on the horizontal axis and the probabilities on the vertical axis.

In a probability distribution, each probability is between 0 and 1, inclusive. Because all possible values of the random variable are included in the probability distribution, the sum of the probabilities is 1.

EXAMPLE

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X

be the number of times per week a newborn baby's crying wakes its mother after midnight. For this example, the values of the random variable are $x = 0, 1, 2, 3, 4, 5$.

In the table, the left column contains all of the possible values of the random variable and the right column, $P(x)$, is the probability that X takes on the corresponding value x . For example, in the first row, the value of the random variable is 0, and the probability the random variable is 0 is $\frac{2}{50}$.

In the context of this example, that means that the probability a newborn baby's crying wakes its mother 0 times per week is $\frac{2}{50}$.

x	$P(x)$
0	$\frac{2}{50}$
1	$\frac{11}{50}$
2	$\frac{23}{50}$
3	$\frac{9}{50}$
4	$\frac{4}{50}$
5	$\frac{1}{50}$

Because X can only take on the values 0, 1, 2, 3, 4, and 5, X is a discrete random variable. Note that each probability is between 0 and 1, and the sum of the probabilities is 1:

$$\frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} = 1$$

TRY IT

Suppose Nancy has classes three days a week. She attends classes three days a week 80\% of the time, two days a week 15\% of the time, one day a week 4\% of the time, and no days 1\% of the time. Suppose one week is randomly selected.

1. Let X be the number of days Nancy _____.
2. X takes on what values?
3. Suppose one week is randomly chosen. Construct a probability distribution table like the one in the example above. The table should have two columns labelled x and $P(x)$. What does the $P(x)$ column sum to?

Click to see Solution

1. Let X be the number of days Nancy attends class per week.
2. 0, 1, 2, and 3.

3.

x	$P(x)$
0	0.01
1	0.04
2	0.15
3	0.80

The $P(x)$ column sums to 1.

EXAMPLE

Jeremiah has basketball practice two days a week. Ninety percent of the time, he attends both practices. Eight percent of the time, he attends one practice. Two percent of the time, he does not attend either practice. What is X , and what values does it take on? Construct the probability distribution for this random variable.

Solution

X is the number of days Jeremiah attends basketball practice per week and takes on the values 0, 1, and 2.

x	$P(x)$
0	0.02
1	0.08
2	0.9



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=105#oembed-1>

Video: “Random Variables and Probability Distributions” by Dr Nic’s Maths and Stats [4:39] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=105#oembed-2>

Video: “Constructing a probability distribution for random variable | Khan Academy” by Khan Academy [6:47] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A company wants to evaluate its attrition rate, in other words, how long new hires stay with the company. Over the years, they have established the following probability distribution. Let X be the number of years a new hire will stay with the company. Let $P(x)$ be the probability that a new hire will stay with the company x years. So far, the company has created the following probability distribution table.

x	$P(x)$
0	0.12
1	0.18
2	0.30
3	0.15
4	
5	0.10
6	0.05

- a. What number goes in the empty cell in the table?
- b. $P(x = 1) = ?$
- c. $P(x \geq 5) = ?$
- d. What does the column “ $P(x)$ ” sum to?

Click to see Answer

- a. 0.1

- b. 0.18
- c. 0.15
- d. 1

2. A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer. Let the random variable be the number of batches of muffins sold in the bakery. Through observation, the baker has established a probability distribution.

x	$P(x)$
1	0.15
2	0.35
3	0.40
4	0.10

- a. What is the probability the baker will sell more than one batch?
- b. What is the probability the baker will sell exactly one batch?

Click to see Answer

- a. 0.85
- b. 0.15

3. Ellen has music practice three days a week. She practices for all of the three days 85% of the time, two days 8% of the time, one day 4% of the time, and no days 3% of the time. One week is selected at random.
- a. Define the random variable X .
 - b. Construct a probability distribution table for the data.

Click to see Answer

- a. The number of times Ellen practices each week.

b.

x	$P(x)$
0	0.03
1	0.04
2	0.08
3	0.85

4. Javier volunteers in community events each month. He does not do more than five events in a month. He attends exactly five events 35\% of the time, four events 25\% of the time, three events 20\% of the time, two events 10\% of the time, one event 5\% of the time, and no events 5\% of the time.
- Define the random variable X .
 - What values does x take on?
 - Construct the probability distribution table.
 - Find the probability that Javier volunteers for less than three events each month.
 - Find the probability that Javier volunteers for at least one event each month.

Click to see Answer

- The number of times Javier volunteers each month.
- 0, 1, 2, 3, 4, 5

c.

x	$P(x)$
0	0.05
1	0.05
2	0.1
3	0.2
4	0.25
5	0.35

- 0.2
- 0.95

5. Suppose that the probability distribution for the number of years it takes to earn a Bachelor of Science degree is given in the following table.

x	$P(x)$
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

- In words, define the random variable X .
- What does it mean that the values zero, one, and two are not included for x in the probability distribution?

Click to see Answer

- The number of years it takes to earn a Bachelor of Science degree.
 - The probabilities associated with those values is 0.
6. A physics professor wants to know what percent of physics majors will spend the next several years doing post-graduate research. He has the following probability distribution.

x	$P(x)$
1	0.35
2	0.20
3	0.15
4	
5	0.10
6	0.05

- Define the random variable X .
- Define $P(x)$, or the probability of x .
- Find the probability that a physics major will do post-graduate research for four years.
- Find the probability that a physics major will do post-graduate research for at most three years.

Click to see Answer

- The number of years a physics major spends doing post-graduate research.
- The probability that a physics major spends exactly x years doing post-graduate research.

- c. 0.15
- d. 0.7

7. A ballet instructor is interested in knowing what percent of each year's class will continue on to the next, so that she can plan what classes to offer. Over the years, she has established the following probability distribution.

x	$P(x)$
1	0.1
2	0.05
3	0.1
4	
5	0.3
6	0.2
7	0.1

- a. In words, define the random variable X .
- b. What number goes in the missing cell in the table?
- c. $P(x < 4) = ?$
- d. What does the column $P(x)$ sum to and why?

Click to see Answer

- a. The number of years a student will study ballet with the teacher.
- b. 0.15
- c. 0.25
- d. 1 because all possible values of the random variable are included in the probability distribution.

8. A theatre group holds a fund-raiser. It sells 100 raffle tickets for \$5 apiece. Suppose you purchase four tickets. The prize is two passes to a Broadway show worth a total of \$150.
- a. What are you interested in here?
 - b. In words, define the random variable X .
 - c. List the values that X may take on.
 - d. Construct the probability distribution.

Click to see Answer

- a. How much money you will win or lose.
- b. The amount of money you spent/earned.
- c. $-20, 130$

d.

x	$P(x)$
-20	0.96
130	0.04

9. Suppose that you are offered the following “deal.” You roll a die. If you roll a six, you win \$10. If you roll a four or five, you win \$5. If you roll a one, two, or three, you pay \$6.
- a. What are you ultimately interested in here (the value of the roll or the money you win)?
 - b. In words, define the random variable X .
 - c. List the values that X may take on.
 - d. Construct the probability distribution.

Click to see Answer

- a. How much money you will win or lose.
- b. The amount of money you win or lose.
- c. $-6, 5, 10$

d.

x	$P(x)$
-6	0.5
5	0.3333
10	0.1667

10. People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given in the following table. There is a five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	$P(x)$
0	0.03
1	0.05
2	0.24
3	
4	0.07
5	0.04

- Describe the random variable X in words.
- Find the probability that a customer rents three DVDs.
- Find the probability that a customer rents at least four DVDs.
- Find the probability that a customer rents at most two DVDs.

Click to see Answer

- The number of DVDs a person rents at any one time.
- 0.57
- 0.11
- 0.32

“4.2 Probability Distribution of a Discrete Random Variable” and “4.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

4.3 EXPECTED VALUE AND STANDARD DEVIATION FOR A DISCRETE PROBABILITY DISTRIBUTION

LEARNING OBJECTIVES

- Calculate and interpret the expected value of a probability distribution.
- Calculate the standard deviation for a probability distribution.

Expected Value of a Probability Distribution

The **expected value** is often referred to as the “**long-term**” **average or mean**. That is, over the long term of repeatedly doing an experiment, the expected outcome is this average.

Suppose we toss a coin and record the result. What is the probability that the result is heads? If we flip a coin two times, does the probability tell us that these flips will result in one head and one tail? We might toss a fair coin ten times and record nine heads. Probability does not describe the short-term results of an experiment. Probability gives information about what can be expected in the long term. To demonstrate this, Karl Pearson once tossed a fair coin 24,000 times! He recorded the results of each toss, obtaining heads 12,012 times, meaning that roughly half of the flips resulted in heads. In his experiment, Pearson illustrated the Law of Large Numbers.

The Law of Large Numbers states that, as the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency approaches zero—the theoretical probability and the relative frequency get closer and closer together. This is exactly what Karl Pearson observed when he tossed the coin 24,000 times. When evaluating the long-term results of statistical experiments, we often want to know the “average”

outcome. This “long-term average” is known as the **mean** or **expected value** of the experiment. In other words, after conducting many trials of an experiment, we would expect this average value.

The **expected value**, denoted by μ or $E(x)$, is a weighted average where each value of the random variable is weighted by the value’s corresponding probability.

$$E(x) = \sum (x \times P(x))$$

EXAMPLE

A men’s soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is 0.2, the probability that they play one day is 0.5, and the probability that they play two days is 0.3. Find the long-term average or expected value of the number of days per week the men’s soccer team plays soccer.

Solution

First, let the random variable X be the number of days the men’s soccer team plays soccer per week. X takes on the values 0, 1, 2. The table below shows the probability distribution for X , and includes an additional column $x \times P(x)$ that we will use to calculate the expected value. In this new column, we will multiply each x value by its corresponding probability.

x	$P(x)$	$x \times P(x)$
0	0.2	$0 \times 0.2 = 0$
1	0.5	$1 \times 0.5 = 0.5$
2	0.3	$2 \times 0.3 = 0.6$

Add the last column to find the long-term average or expected value:

$$\begin{aligned}
 E(x) &= (0 \times 0.2) + (1 \times 0.5) + (2 \times 0.3) \\
 &= 0 + 0.5 + 0.6 \\
 &= 1.1
 \end{aligned}$$

The expected value is 1.1. The men’s soccer team would, on the average, expect to play soccer 1.1

days per week. The number **1.1** is the long-term average or expected value if the men's soccer team plays soccer week after week after week.

NOTE

The expected value does not represent a value that the random variable takes on. The expected value is an average. In this case, the expected value of **1.1** is the average times the team plays per week. To understand what this means, imagine that each week, we record the number of times the soccer team played that week. We do this repeatedly for many, many, many weeks. Then, we calculate the mean of the numbers we recorded (using the techniques we learned previously)—the mean of these numbers equals **1.1**, the expected value. The number of trials must be very, very large in order for the mean of the values recorded from the trials to equal the expected value calculated using the expected value formula.

EXAMPLE

Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from zero to nine with replacement. You pay \$2 to play and could profit \$100,000 if you match all five numbers in order (you get your \$2 back plus \$100,000). Over the long term, what is your **expected** profit from playing the game?

Solution

To solve this problem, set up an expected value table for the amount of money you can profit. Let X be the amount of money you profit. The values of x are not 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Because you are interested in your **profit (or loss)**, the values of x are \$100,000 and —\$2 dollars.

To win, you must get all five numbers correct, in order. Because there are ten numbers, the

probability of choosing one correct number is $\frac{1}{10} = 0.1$. You may choose a number more than once. The probability of choosing all five numbers correctly and in order is

$$\frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} = 0.00001$$

Therefore, the probability of winning is 0.00001 and the probability of losing is $1 - 0.00001 = 0.99999$.

The expected value is as follows:

x	$P(x)$	$x \times P(x)$
-2	0.99999	-1.99998
100,000	0.00001	1

$$\begin{aligned} E(x) &= -1.99998 + 1 \\ &= -0.99998 \end{aligned}$$

Because -0.99998 is about -1 , you would, on average, expect to lose approximately \$1 for each game you play. However, each time you play, you either lose \$2 or profit \$100,000. The \$1 is the average (or expected) LOSS per game after playing this game over and over.

EXAMPLE

Suppose you play a game with a biased coin where the probability of heads is $\frac{2}{3}$. You play each game by tossing the coin once. If you toss a head, you pay \$6. If you toss a tail, you win \$10. If you play this game many times, will you come out ahead?

1. Define a random variable X .
2. Construct the probability distribution for X .

3. What is the expected value? Do you come out ahead?

Solution

1. Let X be the amount of profit per game. The values of x are $-\$6$ (for a loss) and $\$10$ (for a win).

2.

x	$P(x)$
10	$\frac{1}{3}$
-6	$\frac{2}{3}$

3.

x	$P(x)$	$x \times P(x)$
10	$\frac{1}{3}$	$\frac{10}{3}$
-6	$\frac{2}{3}$	-4

$$\begin{aligned} E(x) &= \frac{10}{3} + (-4) \\ &= -0.67 \end{aligned}$$

On average, you lose $\$0.67$ each time you play the game, so you do not come out ahead.

TRY IT

You are playing a game of chance in which four cards are drawn from a standard deck of 52 cards.

You guess the suit of each card before it is drawn. The cards are replaced in the deck on each draw. You pay \$1 to play. If you guess the right suit every time, you get your money back and \$256. What is your expected profit from playing the game over the long term?

Click to see Solution

Let X be the amount of money you profit. The values of x are $-\$1$ (for a loss) and $\$256$ (for a win).

The probability of winning (guessing the correct suit on each draw) is

$$\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = 0.0039$$

The probability of losing is

$$1 - 0.0039 = 0.9961$$

The expected value is as follows:

x	$P(x)$	$x \times P(x)$
-1	0.9961	-0.9961
256	0.0039	0.9984

$$\begin{aligned} E(x) &= -0.9961 + 0.9984 \\ &= 0.0023 \end{aligned}$$

Playing the game over and over again means you would average \$0.0023 in profit per game.

TRY IT

Suppose you play a game with a spinner that has three colours on it: red, green, and blue. The

probability of landing on red is 40%, and the probability of landing on green is 20%. You play a game by spinning the spinner once. If you land on red, you pay \$10. If you land on blue, you do not pay or win anything. If you land on green, you win \$10. What is the expected value of this game? Do you come out ahead?

Click to see Solution

Let X be the amount won in a game. The values of x are $-\$10$ (for red), $\$0$ (for blue) and $\$10$ (for a green).

x	$P(x)$	$x \times P(x)$
-10	0.4	-4
0	0.4	0
10	0.2	2

$$\begin{aligned} E(x) &= -4 + 0 + 2 \\ &= -2 \end{aligned}$$

On average, you lose \$2 per game. So you do not come out ahead.

Standard Deviation of a Probability Distribution

Like data, probability distributions have standard deviations. The **standard deviation**, denoted σ , of a probability distribution for a random variable X describes the spread or variability of the probability distribution. The standard deviation is the standard deviation we expect when doing an experiment over and over.

$$\sigma = \sqrt{\sum ((x - \mu)^2 \times P(x))}$$

To calculate the standard deviation of a probability distribution, find each deviation from its expected value, square it, multiply it by its probability, add the products, and take the square root.

EXAMPLE

Let X be the number of times per week a newborn baby's crying wakes its mother after midnight. The probability distribution for X is:

x	$P(x)$
0	0.04
1	0.22
2	0.46
3	0.18
4	0.08
5	0.02

Find the expected value and standard deviation of the number of times a newborn baby's crying wakes its mother after midnight.

Solution

For the expected value:

x	$P(x)$	$x \times P(x)$
0	0.04	0
1	0.22	0.22
2	0.46	0.92
3	0.18	0.54
4	0.08	0.32
5	0.02	0.1

$$\begin{aligned}\mu &= 0 + 0.22 + 0.92 + 0.54 + 0.32 + 0.1 \\ &= 2.1\end{aligned}$$

On average, a newborn wakes its mother after midnight **2.1** times per week.

For the standard deviation: For each value x , multiply the square of its deviation by its probability (each deviation has the format $x - \mu$).

x	$P(x)$	$(x - \mu)^2 \times P(x)$
0	0.04	$(0 - 2.1)^2 \times 0.04 = 0.1764$
1	0.22	$(1 - 2.1)^2 \times 0.22 = 0.2662$
2	0.46	$(2 - 2.1)^2 \times 0.46 = 0.0046$
3	0.18	$(3 - 2.1)^2 \times 0.18 = 0.1458$
4	0.08	$(4 - 2.1)^2 \times 0.08 = 0.2888$
5	0.02	$(5 - 2.1)^2 \times 0.02 = 0.1682$
Sum		1.05

Add the values in the third column of the table and then take the square root of this sum:

$$\begin{aligned}\sigma &= \sqrt{1.05} \\ &= 1.024 \dots\end{aligned}$$

TRY IT

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. Let X be the number of times a post-op patient rings for the nurse. For a random sample of 50 patients, the following information was obtained. What is the expected value? What is the standard deviation?

x	$P(x)$
0	0.08
1	0.16
2	0.32
3	0.28
4	0.12
5	0.04

Click to see Solution

For the expected value:

x	$P(x)$	$x \times P(x)$
0	0.08	0
1	0.16	0.16
2	0.32	0.64
3	0.28	0.84
4	0.12	0.48
5	0.04	0.2

$$\begin{aligned}\mu &= 0 + 0.16 + 0.64 + 0.84 + 0.48 + 0.2 \\ &= 2.32\end{aligned}$$

For the standard deviation:

x	$P(x)$	$(x - \mu)^2 \times P(x)$
0	0.08	0.430592
1	0.16	0.278784
2	0.32	0.032768
3	0.28	0.129472
4	0.12	0.338688
5	0.04	0.287296

$$\begin{aligned}\sigma &= \sqrt{0.430592 + 0.278784 + 0.032768 + 0.129472 + 0.338688 + 0.287296} \\ &= 1.22 \dots\end{aligned}$$

TRY IT

On May 11, 2013, at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Japan was about 1.08%. You bet that a moderate earthquake will occur in Japan during this period. If you win the bet, you win \$100. If you lose the bet, you pay \$10. Let X be the amount of profit from a bet. Find the mean and standard deviation of X .

Click to see Solution

x	$P(x)$	$x \times P(x)$	$(x - \mu)^2 \times P(x)$
100	0.0108	1.08	127.8726
-10	0.9892	-9.892	1.3961

$$\begin{aligned}\mu &= 1.08 + (-9.892) \\ &= -8.812\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{127.7826 + 1.3961} \\ &= 11.3696 \dots\end{aligned}$$



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=107#oembed-1>

Video: “Mean (expected value) of a discrete random variable | AP Statistics | Khan Academy” by Khan Academy [4:32] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=107#oembed-2>

Video: “Variance and standard deviation of a discrete random variable | AP Statistics | Khan Academy” by Khan Academy [6:26] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A company wants to evaluate its attrition rate, in other words, how long new hires stay with the company. Over the years, they have established the following probability distribution. Let X be the number of years a new hire will stay with the company.

x	$P(x)$
0	0.12
1	0.18
2	0.30
3	0.15
4	0.10
5	0.10
6	0.05

- a. On average, how long would you expect a new hire to stay with the company?
- b. Calculate the standard deviation for the probability distribution.

Click to see Answer

- a. 2.43 years
 - b. 1.65 years
2. A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer. Through observation, the baker has established a probability distribution. The random variable X is the number of batches of muffins the baker sells.

x	$P(x)$
1	0.15
2	0.35
3	0.40
4	0.10

- a. On average, how many batches should the baker make?

- b. Calculate the standard deviation for the probability distribution.

Click to see Answer

- a. 2.45 batches
b. 0.86 batches

3. A physics professor wants to know what percent of physics majors will spend the next several years doing post-graduate research. Let the random variable be the number of years a physics major will spend doing post-graduate research. The professor has the following probability distribution.

x	$P(x)$
1	0.35
2	0.20
3	0.15
4	0.15
5	0.10
6	0.05

- a. On average, how many years would you expect a physics major to spend doing post-graduate research?
b. Calculate the standard deviation for the probability distribution.

Click to see Answer

- a. 2.6 years
b. 1.56 years

4. A ballet instructor is interested in knowing what percent of each year's class will continue on to the next, so that she can plan what classes to offer. Over the years, she has established the following probability distribution. Let the random variable be the number of years a student will study ballet with the teacher.

x	$P(x)$
1	0.10
2	0.05
3	0.10
4	0.15
5	0.30
6	0.20
7	0.10

- On average, how many years would you expect a child to study ballet with this teacher?
- Calculate the standard deviation for the probability distribution.

Click to see Answer

- 4.5 years
 - 1.72 years
5. You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win \$30. If it is not a face card, you pay \$2. There are 12 face cards in a deck of 52 cards. Let the random variable be the amount of money you win/lose for each draw from the deck.
- Construct the probability distribution for this random variable.
 - What is the expected value of playing the game?
 - Should you play the game? Explain.

Click to see Answer

a.

x	$P(x)$
30	$\frac{12}{52}$
-2	$\frac{40}{52}$

- \$5.38
- Yes, because, on average, you will win \$5.38 for each draw.

6. A theatre group holds a fund-raiser. It sells 100 raffle tickets for \$5 apiece. Suppose you purchase four tickets. The prize is two passes to a Broadway show, worth a total of \$150. Let the random variable be the amount of money you spent/earned in the raffle.
- Construct the probability distribution for this random variable.
 - If this fund-raiser is repeated often and you always purchase four tickets, what would be your expected average winnings per raffle?
 - Calculate the standard deviation for the probability distribution.

Click to see Answer

a.

x	$P(x)$
-20	0.96
130	0.04

- \$14
- \$29.39

7. A game involves selecting a card from a regular 52-card deck and tossing a coin. The coin is a fair coin and is equally likely to land on heads or tails.
- If the card is a face card and the coin lands on Heads, you win \$6.
 - If the card is a face card and the coin lands on Tails, you win \$2.
 - If the card is not a face card, you lose \$2, no matter what the coin shows.

Let the random variable be the amount of money you win/lose for each play.

- Construct the probability distribution for this random variable.
- Find the expected value for this game (expected net gain or loss).
- Explain what your calculations indicate about your long-term average profits and losses on this game.
- Should you play this game to win money?

Click to see Answer

a.

x	$P(x)$
6	0.1154
2	0.1154
-2	0.7692

b. $-\$0.62$

c. If you play the game repeatedly a large number of times, you will lose an average of \$0.62 per play.

d. You should not play the game because, on average, you will lose \$0.62 per play.

8. You buy a lottery ticket to a lottery that costs \$10 per ticket. There are only 100 tickets available to be sold in this lottery. In this lottery there are one \$500 prize, two \$100 prizes, and four \$25 prizes. Find your expected gain or loss.

Click to see Answer

\$0

9. Suppose that you are offered the following “deal.” You roll a die. If you roll a six, you win \$10. If you roll a four or five, you win \$5. If you roll a one, two, or three, you pay \$6. Let the random variable be the amount of money you win/lose on each play.
- Construct a probability distribution for this random variable.
 - Over the long run of playing this game, what are your expected average winnings per game?
 - Based on numerical values, should you take the deal? Explain your decision in complete sentences.

Click to see Answer

a.

x	$P(x)$
-6	0.5
5	0.3333
10	0.1667

b. \$0.33 per play.

c. You should play the game because, on average, you will win \$0.33 per play.

10. A venture capitalist willing to invest \$1,000,000 has three investments to choose from.

- The first investment, a software company, has a 10\% chance of returning \$5,000,000 profit, a 30\% chance of returning \$1,000,000 profit, and a 60\% chance of losing the million dollars
- The second investment, a hardware company, has a 20\% chance of returning \$3,000,000 profit, a 40\% chance of returning \$1,000,000 profit, and a 40\% chance of losing the million dollars.
- The third investment, a biotech firm, has a 10\% chance of returning \$6,000,000 profit, a 70\% chance of returning \$1,000,000, and a 20\% chance of losing the million dollars.

Let X be the profit/loss for the first investment, let Y be the profit/loss for the second investment, and let Z be the profit/loss for the third investment.

- a. Construct a probability distribution for the first investment.
- b. Construct a probability distribution for the second investment.
- c. Construct a probability distribution for the third investment.
- d. Find the expected value for each investment.
- e. Which investment has the highest expected return, on average?

Click to see Answer

a.

x	$P(x)$
4,000,000	0.1
0	0.3
-1,000,000	0.6

b.

x	$P(x)$
2,000,000	0.2
0	0.4
-1,000,000	0.4

c.

x	$P(x)$
5,000,000	0.1
0	0.7
-1,000,000	0.2

- d. Investment 1: \$200,000; Investment 2: \$0; Investment 3: \$300,000.
 e. Investment 3.

11. Suppose that the probability distribution for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given in the following table. On average, how many years do you expect it to take for an individual to earn a B.S.?

x	$P(x)$
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

Click to see Answer

4.85 years

12. A “friend” offers you the following “deal.” For a \$10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.
- Ten of the coupons are for a free gift worth \$6.
 - Eighty of the coupons are for a free gift worth \$8.
 - Six of the coupons are for a free gift worth \$12.
 - Four of the coupons are for a free gift worth \$40.

Based on the financial gain or loss over the long run, should you play the game?

Click to see Answer

No, because, on average, you will lose —\$0.68 per play.

“4.3 Expected Value and Standard Deviation for a Discrete Probability Distribution” and “4.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

4.4 THE BINOMIAL DISTRIBUTION

LEARNING OBJECTIVES

- Recognize the binomial probability distribution and apply it appropriately.

A **binomial experiment** has the following characteristics:

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only two possible outcomes, called “success” and “failure,” for each trial. The letter p denotes the probability of a success on any one trial, and $1 - p$ denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.
4. For each individual trial, the probability of a success, p , and the probability of a failure, $1 - p$, remain the same. Because the n trials are independent, the outcome of one trial does not affect the outcome of another trial.

For example, randomly guessing at a true-false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Joe always guesses correctly on any statistics true-false question with probability $p = 0.6$. Then, $1 - p = 0.4$. This means that for every true-false statistics question Joe answers, his probability of success $p = 0.6$ and his probability of failure $1 - p = 0.4$ remain the same.

The outcomes of a binomial experiment fit a **binomial probability distribution**. The random variable X is the number of successes obtained in the n independent trials. The mean of a binomial probability distribution is $\mu = n \times p$, and the standard deviation is $\sigma = \sqrt{n \times p \times (1 - p)}$

Any experiment with the characteristics of a binomial experiment and where $n = 1$ is called a **Bernoulli Trial** (named after Jacob Bernoulli, who, in the late 1600s, studied them extensively). A binomial experiment takes place when the number of successes is counted in one or more Bernoulli Trials.

EXAMPLE

At ABC College, the withdrawal rate from an elementary physics course is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A “success” could be defined as an individual who withdrew from the course. The random variable X is the number of students who withdraw from the randomly selected elementary physics class.

TRY IT

The health board is concerned about the amount of fruit available in school lunches. 48% of schools in the state offer fruit in their lunches every day. This implies that 52% do not. What would a “success” be in this case?

Click to see Solution

A success would be a school that offers fruit in their lunch every day.

EXAMPLE

Suppose a game has only two outcomes: win or lose. The probability of winning any game is 55%, and the probability of losing is 45%. Each game played is independent. Suppose someone plays the game 20 times. Is this a binomial experiment?

Solution

- There are 20 trials (games).
- There are only two possible outcomes in any trial (game): win (success) or lose (failure). The probability of success (winning) is $p = 0.55$. The probability of a failure is $1 - p = 0.45$.
- Each trial (game) is independent. The outcome of any game does not affect the outcome of any other game.
- The probability of success and the probability of failure remain the same from game to game.

EXAMPLE

Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

1. This is a binomial problem because there is only a success or a _____, there are a fixed number of trials, and the probability of a success is 0.70 for each trial.
2. If we are interested in the number of students who do their homework on time, then how do we define X ?
3. What values does x take on?
4. What is a “failure,” in words?

5. What is the probability of “failure”?
6. The words “at least” translate to what kind of inequality for the probability question $P(x \geq 40)$.

Solution

1. failure
2. X is the number of statistics students who do their homework on time.
3. $0, 1, 2, \dots, 50$.
4. Failure is defined as a student who does not complete his or her homework on time.
5. $1 - p = 0.30$
6. “At least” means greater than or equal to (\geq). The probability question is $P(x \geq 40)$.

TRY IT

Sixty-five percent of people pass the state driver’s exam on the first try. A group of 50 individuals who have taken the driver’s exam is randomly selected. Why this is a binomial problem?

Click to see Solution

- There are only two outcomes on any exam (pass or fail).
- There is a fixed number of trials ($n = 50$).
- The probability of pass (65%) is the same for each trial.
- The trials are independent. (The fact that any one person passes or fails the exam does not affect whether or not any other person passes or fails.)

EXAMPLE

The following example illustrates a problem that is not binomial. It violates the condition of independence. ABC College has a student advisory committee made up of ten staff members and six students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students?

Solution

The names of all committee members are put into a box, and two names are drawn without replacement. The first name drawn determines the chairperson, and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is $\frac{6}{16}$, and the probability of a student on the second draw is $\frac{5}{15}$. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

TRY IT

A high school lacrosse team is selecting a captain. The names of all the seniors are put into a hat and the first three that are drawn will be the captains. The names are not replaced once they are drawn (one person cannot be two captains). We want to see if the captains all play the same position. State whether or not this is binomial and state why.

Click to see Solution

This is not binomial because the names are not replaced after each draw, which means the probability changes for each time a name is drawn. This violates the condition of independence.

Calculating Binomial Probabilities

CALCULATING BINOMIAL PROBABILITIES IN EXCEL

To calculate probabilities associated with binomial random variables in Excel, use the **binom.dist(x,n,p,logic operator)** function.

- For **x**, enter the number of successes.
- For **n**, enter the number of trials.
- For **p**, enter the probability of success.
- For the logic operator, enter **false** to find the probability of exactly **x** successes and enter **true** to find the probability of at most (less than or equal to) **x** successes.

The output from the **binom.dist** function is:

- The probability of getting exactly **x** success in **n** trials with a probability of success **p** when the logic operator is **false**.
- The probability of at most **x** successes in **n** trials with a probability of success **p** when the logic operator is **true**.

Visit the Microsoft page for more information about the **binom.dist** function.

NOTE

Because we can only enter false or true into the logic operator, the **binom.dist** function can only

directly calculate the probability of getting exactly x successes in n trials or getting at most x success in n trials. In order to calculate other binomial probabilities, such as fewer than x successes, more than x successes or at least x successes, we need to manipulate how we use the **binom.dist** function by changing what we enter into the **binom.dist** function, using the complement rule, or both.

EXAMPLE

It has been stated that about 41\% of adult workers have a high school diploma but do not pursue any further education. Suppose 20 adult workers are randomly selected.

1. How many adult workers in the sample are expected to have a high school diploma but do not pursue any further education?
2. What is the probability that exactly 8 of the workers in the sample have a high school diploma but do not pursue further education?
3. What is the probability that at most 12 of the workers in the sample have a high school diploma but do not pursue further education?

Solution

Let X be the number of workers in the sample who have a high school diploma but do not pursue further education. The number of trials is $n = 20$, and the probability of success is $p = 0.41$.

1. $\mu = n \times p = 20 \times 0.41 = 8.2$. On average, in any sample of 20 workers, 8.2 have a high school diploma but do not pursue further education.
2. We want to find $P(x = 8)$.

Function	binom.dist
Field 1	8
Field 2	20
Field 3	0.41
Field 4	false
Answer	0.1790

The probability that exactly 8 of the workers in the sample have a high school diploma but do not pursue further education is 17.9\%.

3. We want to find $P(x \leq 12)$.

Function	binom.dist
Field 1	12
Field 2	20
Field 3	0.41
Field 4	true
Answer	0.9738

The probability that at most 12 of the workers in the sample have a high school diploma but do not pursue further education is 97.38\%.

TRY IT

About 32\% of students participate in a community volunteer program outside of school. Suppose 30 students are selected at random.

1. What is the expected number of students in the sample that participate in a community volunteer program?
2. What is the probability that exactly 10 of the students in the sample participate in a community volunteer program?
3. What is the probability that at most 14 of the students in the sample participate in a community volunteer program?

Click to see Solution

1. $\mu = n \times p = 30 \times 0.32 = 9.6$

2.

Function	binom.dist
Field 1	10
Field 2	30
Field 3	0.32
Field 4	false
Answer	0.1512

3.

Function	binom.dist
Field 1	14
Field 2	30
Field 3	0.32
Field 4	true
Answer	0.9695

EXAMPLE

In the 2013 *Jerry's Artarama* art supplies catalogue, there are 560 pages and 1.5% of the pages feature signature artists. Suppose 100 pages are randomly selected from the catalogue.

1. What is the probability that fewer than 3 of the pages in the sample feature signature artists?
2. What is the probability that more than 5 of the pages in the sample feature signature artists?
3. What is the probability that at least 4 of the pages in the sample feature signature artists?
4. What is the probability that between 2 and 6 of the pages in the sample feature signature artists?

Solution

1. We want to find $P(x < 3)$. We cannot find this probability directly in Excel because the binom.dist function can only calculate $=$ or \leq probabilities. Because x must be an integer (it is the number of pages), $x < 3$ is the same as $x \leq 2$ (of course, in general, this is not true). So $P(x < 3) = P(x \leq 2)$ and $P(x \leq 2)$ is a probability we can calculate with the binom.dist function.

Function	binom.dist
Field 1	2
Field 2	100
Field 3	0.015
Field 4	true
Answer	0.8098

2. We want to find $P(x > 5)$. We cannot find this probability directly in Excel because the binom.dist function can only calculate $=$ or \leq probabilities. The complement of $>$ is \leq , so $P(x > 5) = 1 - P(x \leq 5)$ and $P(x \leq 5)$ is a probability we can calculate with the binom.dist function.

Function	1-binom.dist
Field 1	5
Field 2	100
Field 3	0.015
Field 4	true
Answer	0.0177

3. We want to find $P(x \geq 4)$. We cannot find this probability directly in Excel because the binom.dist function can only calculate $=$ or \leq probabilities. The complement of \geq is $<$, so $P(x \geq 4) = 1 - P(x < 4)$. Because x must be an integer (it is the number of pages), $x < 4$ is the same as $x \leq 3$. So $P(x \geq 4) = 1 - P(x < 4) = 1 - P(x \leq 3)$ and $P(x \leq 3)$ is a probability we can calculate with the binom.dist function.

Function	1-binom.dist
Field 1	3
Field 2	100
Field 3	0.015
Field 4	true
Answer	0.0642

4. We want to find $P(2 \leq x \leq 6)$. We cannot find this probability directly in Excel because the binom.dist function can only calculate $=$ or \leq probabilities. But, $P(2 \leq x \leq 6) = P(x \leq 6) - P(x \leq 1)$. So we can calculate $P(2 \leq x \leq 6)$ as the difference of two binom.dist functions.

Function	binom.dist	-binom.dist
Field 1	6	1
Field 2	100	100
Field 3	0.015	0.015
Field 4	true	true
Answer	0.4426	

TRY IT

According to a Gallup poll, 60\% of American adults prefer saving over spending. Suppose 50 American adults are selected at random.

1. What is the probability that at least 35 adults in the sample prefer saving over spending?
2. What is the probability that fewer than 20 adults in the sample prefer saving over spending?
3. What is the probability between 15 and 25 adults in the sample prefer saving over spending?
4. What is the probability that more than 30 adults prefer saving over spending?

Click to see Solution

1.	Function	1-binom.dist
	Field 1	34
	Field 2	50
	Field 3	0.6
	Field 4	true
	Answer	0.0955

2.	Function	binom.dist
	Field 1	19
	Field 2	50
	Field 3	0.6
	Field 4	true
	Answer	0.0014

3.

Function	binom.dist	-binom.dist
Field 1	25	14
Field 2	50	50
Field 3	0.6	0.6
Field 4	true	true
Answer	0.0978	

4.

Function	1-binom.dist
Field 1	30
Field 2	50
Field 3	0.6
Field 4	true
Answer	0.4465

TRY IT

During the 2013 regular NBA season, DeAndre Jordan of the Los Angeles Clippers had the highest field goal completion rate in the league. DeAndre scored with 61.3% of his shots. Suppose we take a random sample of 80 shots made by DeAndre during the 2013 season.

1. What is the expected number of shots that scored points in a sample of 80 of DeAndre's shots?
2. What is the probability that DeAndre scored on 60 of the 80 shots?
3. What is the probability that DeAndre scored on more than 50 of the 80 shots?
4. What is the probability that DeAndre scored between 65 and 75 of the 80 shots?

Click to see Solution

1. $\mu = n \times p = 80 \times 0.613 = 49.04$

2.

Function	binom.dist
Field 1	60
Field 2	80
Field 3	0.613
Field 4	false
Answer	0.0036

3.

Function	1-binom.dist
Field 1	50
Field 2	80
Field 3	0.613
Field 4	true
Answer	0.3718

4.

Function	binom.dist	-binom.dist
Field 1	75	64
Field 2	80	80
Field 3	0.613	0.613
Field 4	true	true
Answer	0.0001	



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=109#oembed-1>

Video: “Binomial Probability in Excel – Word problems” by Joshua Emmanuel [7:00] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he or she truly has the flu (and not just a nasty cold) is only about 4%. Suppose a sample of 25 patient calls is taken.
 - a. On average, for every 25 patient calls, how many do you expect to have the flu?
 - b. Calculate the standard deviation for this probability distribution.
 - c. Find the probability that exactly 5 of the 25 patients who called have the flu.
 - d. Find the probability that at most 3 of the 25 patients who called have the flu.
 - e. Find the probability that between 3 and 9 of the patients who called have the flu.
 - f. Find the probability that at least 6 of the 25 patients who called have the flu.
 - g. Find the probability that fewer than 5 of the 25 patients who called have the flu.
 - h. Find the probability that more than 4 of the 25 patients who called have the flu.

Click to see Answer

- a. 1
 - b. 0.98
 - c. 0.0024
 - d. 0.9835
 - e. 0.0765
 - f. 0.0004
 - g. 0.9972
 - h. 0.0165

2. The probability that the San Jose Sharks will win any given game is 0.3694 based on a 13-year win history of 382 wins out of 1,034 games played (as of a certain date). An upcoming monthly schedule contains 12 games.
 - a. What is the expected number of wins for that upcoming month?
 - b. What is the probability that the San Jose Sharks win exactly 6 games in that upcoming month?
 - c. What is the probability that the San Jose Sharks win at least 5 games in that upcoming month?

- d. What is the probability that the San Jose Sharks win between 3 and 9 games in that upcoming month?
- e. What is the probability that the San Jose Sharks win fewer than 6 games in that upcoming month?
- f. What is the probability that the San Jose Sharks win more than 8 games in that upcoming month?
- g. What is the probability that the San Jose Sharks win at most 7 games in that upcoming month?

Click to see Answer

- a. 4.4328
- b. 0.1476
- c. 0.4734
- d. 0.7024
- e. 0.7427
- f. 0.0084
- g. 0.9644

3. A student takes a ten-question true-false quiz but did not study and randomly guesses each answer. Find the probability that the student passes the quiz with a grade of at least 70\% of the questions correct.

Click to see Answer

0.1719

4. More than 96\% of the very largest colleges and universities (more than 15, 000 total enrollments) have some online offerings. Suppose you randomly select 13 such institutions.
- a. On average, how many schools would you expect to offer such courses?
 - b. Find the probability that at most 10 offer such courses.
 - c. Find the probability that more than 9 offer such courses.
 - d. Find the probability that exactly 11 offer such courses.
 - e. Find the probability that between 7 and 12 offer such courses.
 - f. Is it more likely that 12 or that 13 will offer such courses? Use numbers to justify your answer numerically and answer in a complete sentence.

Click to see Answer

- a. 12.48
- b. 0.0135

- c. 0.9986
- d. 0.0797
- e. 0.4118
- f. 13 is more likely because the probability that 13 offer such courses is 0.5882 and the probability that 12 offer such courses is 0.3186.

5. Suppose that about 85\% of graduating students attend their graduation. A group of 22 graduating students is randomly chosen.
- a. How many are expected to attend their graduation?
 - b. Find the probability that between 16 and 18 attend graduation.
 - c. Find the probability that fewer than 15 attend graduation.
 - d. Find the probability that at least 20 attend graduation.
 - e. Would you be surprised if all 22 attended graduation? Justify your answer numerically.

Click to see Answer

- a. 18.7
- b. 0.3880
- c. 0.0114
- d. 0.3382
- e. Yes, because the probability that all 22 attend is only 0.028.

6. At The Fencing Center, 60\% of the fencers use the foil as their main weapon. We randomly survey 25 fencers at The Fencing Center.
- a. How many of the 25 fencers are expected to use the foil as their main weapon?
 - b. Find the probability that 6 of the 25 fencers use the foil as their main weapon.
 - c. Find the probability that at least 15 of the 25 fencers use the foil as their main weapon.
 - d. Find the probability that at most 10 of the 25 fencers use the foil as their main weapon.
 - e. Find the probability that between 12 and 17 of the 25 fencers use the foil as their main weapon.
 - f. Find the probability that fewer than 13 of the 25 fencers use the foil as their main weapon.
 - g. Find the probability that more than 20 of the 25 fencers use the foil as their main weapon.
 - h. Would you be surprised if all 25 fencers in the sample use the foil as their main weapon? Justify your answer numerically.

Click to see Answer

- a. 15

- b. 0.0002
- c. 0.4246
- d. 0.0344
- e. 0.7686
- f. 0.1538
- g. 0.0095
- h. Yes, because the probability that all 25 use the foil is only 0.0000028.

7. Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A sample of 60 seniors is randomly chosen. Of interest is the number who participated in after-school sports all four years of high school.
- a. How many seniors in the sample are expected to have participated in after-school sports all four years of high school?
 - b. What is the probability that at least 10 of the seniors in the sample have participated in after-school sports all four years of high school?
 - c. What is the probability that between 7 and 15 of the seniors in the sample have participated in after-school sports all four years of high school?
 - d. What is the probability that less than 6 of the seniors in the sample have participated in after-school sports all four years of high school?
 - e. Would you be surprised if none of the seniors in the sample participated in after-school sports all four years of high school? Justify your answer numerically.
 - f. Is it more likely that 4 or that 5 of the seniors in the sample participated in after-school sports all four years of high school? Justify your answer numerically.

Click to see Answer

- a. 4.8
 - b. 0.02
 - c. 0.202
 - d. 0.6526
 - e. Yes, because the probability that 0 of the 60 seniors in the sample participated in after-school sports all four years of high school is only 0.0067.
 - f. 4 is more likely because the probability that 4 of the seniors in the sample participated in after-school sports all four years of high school is 0.1873 and the probability that 5 of the seniors in the sample participated in after-school sports all four years of high school is 0.1824.
8. It has been estimated that only about 30% of British Columbia residents have adequate

earthquake supplies. Suppose 11 BC residents are randomly surveyed about their earthquake supplies.

- a. What is the probability that at least 8 of the residents in the sample have adequate earthquake supplies?
- b. What is the probability that between 5 and 9 of the residents in the sample have adequate earthquake supplies?
- c. What is the probability that at most 3 of the residents in the sample have adequate earthquake supplies?
- d. Is it more likely that none or that all of the residents surveyed will have adequate earthquake supplies? Why?
- e. How many residents do you expect will have adequate earthquake supplies?

Click to see Answer

- a. 0.0043
- b. 0.2103
- c. 0.5696
- d. None is more likely because the probability that 0 residents in the sample have adequate earthquake supplies is 0.0198, and the probability that 11 residents in the sample have adequate earthquake supplies is 0.0000018.
- e. 3.3

“4.4 The Binomial Distribution” and “4.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

4.5 THE POISSON DISTRIBUTION

LEARNING OBJECTIVES

- Recognize the Poisson probability distribution and apply it appropriately.

A **Poisson experiment** has the following characteristics:

1. The Poisson probability distribution gives the probability of a number of events occurring in a **fixed interval** of time or space if these events happen with a known average rate and independently of the time since the last event. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on average, there are five words spelled incorrectly in 100 pages. The interval is the 100 pages.
2. The Poisson distribution may be used to approximate the binomial distribution if the probability of success is “small” (such as 0.01) and the number of trials is “large” (such as 1,000).

The random variable X associated with a Poisson experiment is the number of occurrences in the interval of interest. In a Poisson distribution, λ is the average number of occurrences in an interval. The mean of a Poisson probability distribution is $\mu = \lambda$ and the standard deviation is $\sigma = \sqrt{\lambda}$.

EXAMPLE

The average number of loaves of bread put on a shelf in a bakery in a half-hour period is **12**. Of interest is the number of loaves of bread put on the shelf in five minutes. The time interval of interest is five minutes. We want to find the probability that three loaves are put on the shelf in any five minute interval. Why is this a Poisson experiment?

Solution

Let X be the number of loaves of bread put on the shelf in five minutes. If the average number of loaves put on the shelf in 30 minutes (half an hour) is **12**, then the average number of loaves put on the shelf in five minutes is $\frac{5}{30} \times 12 = 2$ loaves of bread.

This is a Poisson experiment because we are interested in the number of loaves happening in a fixed interval (five minutes) with an average of **2** loaves in any five minutes.

Calculating Poisson Probabilities

CALCULATING POISSON PROBABILITIES IN EXCEL

To calculate probabilities associated with a Poisson experiment in Excel, use the **Poisson.dist(x, λ, logic operator)** function.

- For **x**, enter the number of successes over the interval.
- For **λ**, enter the average number of successes over the interval.
- For the logic operator, enter **false** to find the probability of exactly **x** successes and enter **true** to find the probability of at most (less than or equal to) **x** successes.

The output from the **Poisson.dist** function is:

- the probability of getting exactly **x** successes over the interval when the logic operator is **false**.
- the probability of at most **x** successes over the interval when the logic operator is **true**.

Visit the Microsoft page for more information about the **Poisson.dist** function.

NOTE

Because we can only enter false or true into the logic operator, the **Poisson.dist** function can only directly calculate the probability of getting exactly **x** successes or getting at most **x** success over the interval. In order to calculate other Poisson probabilities, such as fewer than **x** successes, more than **x** successes or at least **x** successes, we need to manipulate how we use the **Poisson.dist** function by changing what we enter into the **Poisson.dist** function, using the complement rule, or both.

EXAMPLE

Leah receives about six telephone calls every two hours.

1. What is the probability that Leah receives exactly 4 calls in the next two hours?
2. What is the probability that Leah receives at most 9 calls in the next two hours?
3. What is the probability that Leah receives at most 2 calls in the next hour?

Solution

1. The average number of calls in any two-hour period is 6, so $\lambda = 6$.

Function	Poisson.dist
Field 1	4
Field 2	6
Field 3	false
Answer	0.1339

The probability that Leah receives 4 calls in the next two hours is 13.39\%.

2. The average number of calls in any two-hour period is 6, so $\lambda = 6$.

Function	Poisson.dist
Field 1	9
Field 2	6
Field 3	true
Answer	0.9161

The probability that Leah receives at most 6 calls in the next two hours is 91.61\%.

3. The average number of calls in any two-hour period is 6. So the average number of calls in one hour is $\frac{6}{2} = 3$.

Function	Poisson.dist
Field 1	2
Field 2	3
Field 3	true
Answer	0.4232

The probability that Leah receives at most 6 calls in the next two hours is 42.32\%.

TRY IT

The customer service department of a technology company receives an average of 10 phone calls every hour.

1. What is the probability that the customer service department receives exactly 7 phone calls in an hour?
2. What is the probability that the customer service department receives exactly 2 phone calls in a 15 minute period?
3. What is the probability that the customer service department receives at most 4 phone calls in a 30 minute period?
4. What is the probability that the customer service department receives at most 20 phone calls in a three-hour period?

Click to see Solution

1.	Function	Poisson.dist
	Field 1	7
	Field 2	10
	Field 3	false
	Answer	0.0901

2.	Function	Poisson.dist
	Field 1	2
	Field 2	2.5
	Field 3	false
	Answer	0.2565

3.

Function	Poisson.dist
Field 1	4
Field 2	5
Field 3	true
Answer	0.4405

4.

Function	Poisson.dist
Field 1	20
Field 2	30
Field 3	true
Answer	0.0353

EXAMPLE

According to Baydin, an email management company, an email user gets, on average, **147** emails over a six-hour period.

1. What is the probability that an email user receives fewer than **160** emails over a six-hour period?
2. What is the probability that an email user receives more than **40** emails over a two-hour period?
3. What is the probability that an email user receives at least **600** emails over a **24** hour period?
4. What is the probability that an email user receives between **150** and **200** emails over a six-hour period?

Solution

1. The average over a six-hour period is **147**. We want to find $P(x < 160)$. We cannot find

this probability directly in Excel because the Poisson.dist function can only calculate $=$ or \leq probabilities. Because x must be an integer (it is the number of emails), $x < 160$ is the same as $x \leq 159$. So $P(x < 160) = P(x \leq 159)$ and $P(x \leq 159)$ is a probability we can calculate with the Poisson.dist function.

Function	Poisson.dist
Field 1	159
Field 2	147
Field 3	true
Answer	0.8486

The probability a user receives fewer than 160 emails over a six-hour period is 84.86\%.

2. The average over a two-hour period is $\frac{147}{3} = 49$. We want to find $P(x > 40)$. We cannot find this probability directly in Excel because the Poisson.dist function can only calculate $=$ or \leq probabilities. The complement of $>$ is \leq , so $P(x > 40) = 1 - P(x \leq 40)$ and $P(x \leq 40)$ is a probability we can calculate with the Poisson.dist function.

Function	1-Poisson.dist
Field 1	40
Field 2	49
Field 3	true
Answer	0.8902

The probability a user receives more than 40 emails over a two-hour period is 89.02\%.

3. The average over a 24 hour period is $147 \times 4 = 588$. We want to find $P(x \geq 600)$. We cannot find this probability directly in Excel because the Poisson.dist function can only calculate $=$ or \leq probabilities. The complement of \geq is $<$, so $P(x \geq 600) = 1 - P(x < 600)$. Because x must be an integer (it is the number of emails), $x < 600$ is the same as $x \leq 599$. So $P(x \geq 600) = 1 - P(x < 600) = 1 - P(x \leq 599)$ and $P(x \leq 599)$ is a probability we can calculate with the Poisson.dist function.

Function	1-Poisson.dist
Field 1	599
Field 2	588
Field 3	true
Answer	0.3158

The probability a user receives at least 600 emails over a 24-hour period is 31.58\%.

4. We want to find $P(150 \leq x \leq 200)$. We cannot find this probability directly in Excel because the Poisson.dist function can only calculate $=$ or \leq probabilities. But, $P(150 \leq x \leq 200) = P(x \leq 200) - P(x \leq 149)$. So we can calculate $P(150 \leq x \leq 200)$ as the difference of two Poisson.dist functions.

Function	Poisson.dist	-Poisson.dist
Field 1	200	149
Field 2	147	147
Field 3	true	true
Answer	0.4132	

The probability a user receives between 150 and 200 emails over a six-hour period is 41.32%.

TRY IT

A car parts manufacturer can produce an average of 25 parts from 100 metres of sheet metal.

1. What is the probability that more than 30 parts can be made from 100 metres of sheet metal?
2. What is the probability that between 10 and 20 parts can be made from 50 metres of sheet metal?
3. What is the probability that fewer than 5 parts can be made from 25 metres of sheet metal?

4. What is the probability that at least 80 parts can be made from 400 metres of sheet metal?

Click to see Solution

1.

Function	1-Poisson.dist
Field 1	30
Field 2	25
Field 3	true
Answer	0.1367

2.

Function	Poisson.dist	-Poisson.dist
Field 1	20	9
Field 2	12.5	12.5
Field 3	true	true
Answer	0.7813	

3.

Function	Poisson.dist
Field 1	4
Field 2	6.25
Field 3	true
Answer	0.2530

4.

Function	1-Poisson.dist
Field 1	79
Field 2	100
Field 3	true
Answer	0.9825



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=111#oembed-1>

Video: “The Poisson Distribution – explained with examples and illustrated using Excel – statistics Help” by Dr Nic’s Maths and Stats [7:49] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. On average, a clothing store gets 120 customers per day. Assume the store is open 12 hours each day.
 - a. What is the probability that the store gets 150 customers in one day?
 - b. What is the probability that the store gets 35 customers in a 4-hour period?
 - c. What is the probability that the store gets more than 12 customers in a 1-hour period?
 - d. What is the probability that the store gets fewer than 12 customers in a 2-hour period?
 - e. What is the probability that the store gets between 100 and 130 customers in one day?
 - f. What is the probability that the store gets at most than 20 customers in a 3-hour period?
 - g. What is the probability that the store gets at least 70 customers in a 6-hour period?

Click to see Answer

- a. 0.001
 - b. 0.0485
 - c. 0.2084
 - d. 0.0214
 - e. 0.8036
 - f. 0.0353
 - g. 0.1118

2. On average, 8 teens in the U.S. die from motor vehicle injuries in a 24-hour period.
 - a. What is the probability that at most 3 teens die from motor vehicle injuries in a 24-hour period?
 - b. What is the probability that between 8 and 12 teens die from motor vehicle injuries in a

48-hour period?

- c. What is the probability that at least 3 teens die from motor vehicle injuries in a 6-hour period?
- d. What is the probability that fewer than 7 teens die from motor vehicle injuries in a 12-hour period?
- e. Is it likely that there will be no teens killed from motor vehicle injuries in any 24-hour period in the U.S.? Justify your answer numerically.
- f. Is it likely that there will be more than 20 teens killed from motor vehicle injuries in any 24-hour period in the U.S.? Justify your answer numerically.

Click to see Answer

- a. 0.0424
- b. 0.1711
- c. 0.3233
- d. 0.8893
- e. Not likely because the probability that 0 teens die from motor vehicle injuries in any 24-hour period is 0.0003.
- f. Not likely because the probability that more than 20 teens die from motor vehicle injuries in any 24-hour period is 0.000094.

3. The switchboard in a law office receives an average of 6 calls in an hour.

- a. What is the probability that the office receives exactly 4 calls in a 30-minute period?
- b. What is the probability that the office receives at most 30 calls in a 4-hour period?
- c. What is the probability that the office receives at least 3 calls in a 20-minute period?
- d. What is the probability that the office receives less than 45 calls in an 8-hour period?
- e. What is the probability that the office receives more than 10 calls in a 1-hour period?
- f. What is the probability that the office receives between 15 and 20 calls in a 2-hour period?

Click to see Answer

- a. 0.168
- b. 0.9041
- c. 0.3233
- d. 0.3131
- e. 0.0426
- f. 0.2164

4. The maternity ward at Dr. Jose Fabella Memorial Hospital in Manila, in the Philippines, is

one of the busiest in the world, with an average of 60 births in a 24-hour period.

- a. What is the probability that the maternity ward will deliver 3 babies in a 1-hour period?
- b. What is the probability that the maternity ward will deliver more than 70 babies in a 24-hour period?
- c. What is the probability that the maternity ward will deliver at most 130 babies in a 48-hour period?
- d. What is the probability that the maternity ward will deliver between 10 and 20 babies in a 6-hour period?
- e. What is the probability that the maternity ward will deliver at least 200 babies in a 72-hour period?
- f. What is the probability that the maternity ward will deliver fewer than 8 babies in a 4-hour period?

Click to see Answer

- a. 0.2138
- b. 0.0902
- c. 0.8315
- d. 0.8472
- e. 0.0749
- f. 0.2202

5. Fertile female cats produce an average of 3 litters per year. Suppose that one fertile female cat is randomly chosen.

- a. Find the probability that the cat has no litters in one year.
- b. Find the probability that the cat has at least two litters in one year.
- c. Find the probability that the cat has exactly three litters in one year.

Click to see Answer

- a. 0.0498
- b. 0.8009
- c. 0.224

6. On average, Pierre, an amateur chef, drops 3 pieces of eggshell into every two cake batters he makes.

- a. What is the probability that there will be more than 5 pieces of eggshell in any two cake batters?
- b. What is the probability that there will be at most 20 pieces of eggshell in a dozen cake batters?

- c. What is the probability that there will not be any pieces of eggshell in any single cake batter?
- d. What is the probability that there will be fewer than 10 pieces of eggshell in any four cake batter?
- e. What is the probability that there will be at least 15 pieces of eggshell in any six-cake batter?
- f. What is the probability that there will be between 2 and 4 pieces of eggshell in any two cake batter?
- g. Is it possible for there to be 7 pieces of eggshell in any single cake batter? Why?

Click to see Answer

- a. 0.0839
- b. 0.9884
- c. 0.2231
- d. 0.9161
- e. 0.0415
- f. 0.6161
- g. It is possible but unlikely because the probability of 7 pieces of eggshell in any single cake batter is 0.0008.

7. On a particular nature trail in a national park, deer are spotted at a rate of 1 deer every 3 kilometres.
- a. What is the probability that exactly 4 deer are spotted in any 6 kilometre stretch of trail?
 - b. What is the probability that more than 2 deer are spotted in any 1.5 kilometre stretch of trail?
 - c. What is the probability that between 6 and 10 deer are spotted in any 9-kilometre stretch of trail?
 - d. What is the probability that at most 3 deer are spotted in any 6 kilometre stretch of trail?
 - e. What is the probability that at least 5 deer are spotted in any 3 kilometre stretch of trail?
 - f. What is the probability that fewer than 7 deer are spotted in any 9 kilometre stretch of trail?
 - g. If someone walks the entire 12 kilometres of the trail, is it likely that they will not see any deer? Why?

Click to see Answer

- a. 0.0902
- b. 0.0144
- c. 0.0836

- d. 0.8571
- e. 0.0037
- f. 0.9665
- g. It is unlikely they will not see any deer because the probability of seeing 0 deer in any 12 kilometres is 0.0183.

“4.5 The Poisson Distribution” and “4.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

PART V

CONTINUOUS PROBABILITY DISTRIBUTIONS AND THE NORMAL DISTRIBUTION

Previously, we learned that a random variable is a numerical description of an experiment and that each possible outcome of the experiment is associated with a value of the random variable. In conjunction with the random variable, we can construct the probability distribution of the random variable, which lists all possible values of the random variable along with the probability that the random variable takes on that particular value. In the preceding chapter, we looked at discrete random variables and their associated probability distributions. Because a discrete random variable only takes on certain numerical values, the probability distribution of a discrete random variable is often presented as a table, listing the values of the random variable and their corresponding probabilities.

But what about the probability distribution of a continuous random variable? Recall that a continuous random variable takes on any numerical value in an interval or collection of intervals. Most often, continuous random variables are associated with things that are measured, such as height, weight, temperature, and volume. Continuous random variables have many applications, such as baseball batting averages, IQ scores, the length of time a long-distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores.

As with discrete random variables, a continuous random variable associates a number with the outcome of an experiment, and we are interested in the probability corresponding to the values of the random variable. But unlike a discrete random variable, we cannot make a list of all of the possible values of a continuous random variable. For example, suppose we define a random variable X to be the volume, in litres, of milk in a one-litre milk carton. In this case, the random variable is continuous because the random variable is measuring volume. The random variable X can take on **any** number between 0 and 1. Because there are an infinite number of numbers between 0 and 1, we simply cannot write them all down. This is true for any continuous random variable—it is impossible to write down all of the possible values associated with the random variable. Consequently, we need to look at and work with the probability distribution for a continuous random variable in a different way than we did with discrete random variables. For a continuous random variable, the probability distribution is most often represented by a graph, and

the probabilities associated with the continuous random variable are the corresponding areas under the curve.

CHAPTER OUTLINE

5.1 Probability Distribution of a Continuous Random Variable

5.2 The Normal Distribution

5.3 The Standard Normal Distribution

5.4 Calculating Probabilities for a Normal Distribution

“5.1 Introduction to Continuous Random Variables” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

5.1 PROBABILITY DISTRIBUTION OF A CONTINUOUS RANDOM VARIABLE

LEARNING OBJECTIONS

- Recognize and understand continuous probability distributions.

Previously, we learned about random variables. A **continuous random variable** corresponds to data that can be measured. Examples of continuous random variables include baseball batting averages, IQ scores, the length of time a long-distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores. Because the data is measured, there are an infinite number of numbers that a continuous random variable can take on, and so it is impossible to write down all of the numbers associated with a continuous random variable. Consequently, we represent the probability distribution of a continuous random variable with a graph and calculate probabilities associated with the continuous random variable by finding the corresponding area under the graph.

Continuous Probability Distributions

The graph of a probability distribution for a continuous random variable is a curve. The graph is defined so that the area between the curve and the x -axis is equal to the probability. In other words, the probability a continuous random variable takes on a value in a specific interval is the **area** under the curve of the probability distribution of the continuous random variable.

Properties of a probability distribution for a continuous random variable include:

- The entire area under the curve of the distribution and above the x -axis is equal to 1.

- The probability that the continuous random variable takes on a value in between c and d is the area under the curve of the distribution in between $x = c$ and $x = d$.
- The probability that the continuous random variable exactly equals a particular number (i.e. the probability that the random variable takes on a specific number c) is 0. (The area below the curve and above the x -axis at the single point $x = c$ is a rectangle with height but no width. Therefore the area under the curve at $x = c$ is 0, and so the probability is also 0,

Generally, manual calculations to find the area under the curve of a continuous probability distribution is difficult, and most calculations require the use of computers. We will use the built-in functions in Excel to calculate the area under the continuous probability distribution functions. There are many different continuous probability distributions, including the uniform distribution and the exponential distribution. We will focus on the most important continuous probability distribution—the normal distribution.

NOTE

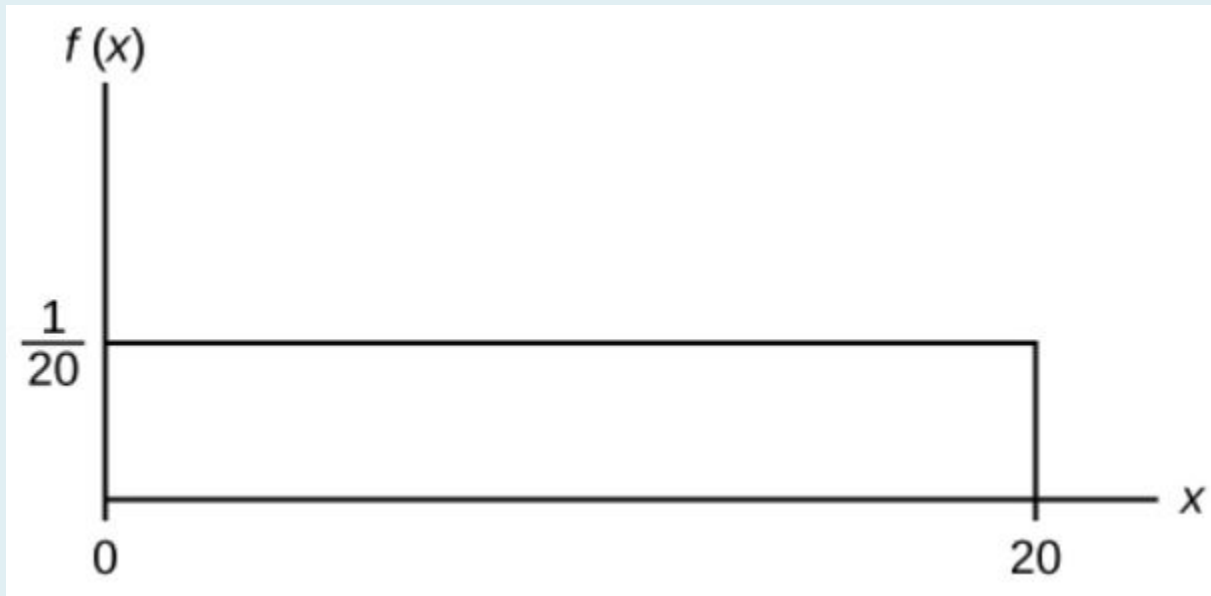
For a continuous random variable, the curve of the probability distribution is denoted by the function $f(x)$. The function $f(x)$ is called a **probability density function**, and $f(x)$ produces the curve of the distribution. The function $f(x)$ is defined so that the **area** between it and the x -axis is equal to a probability. The probability density function $f(x)$ itself does NOT give us probabilities associated with the continuous random variable. The function $f(x)$ produces the graph of the distribution, and the area under this graph corresponds to the probability.

EXAMPLE

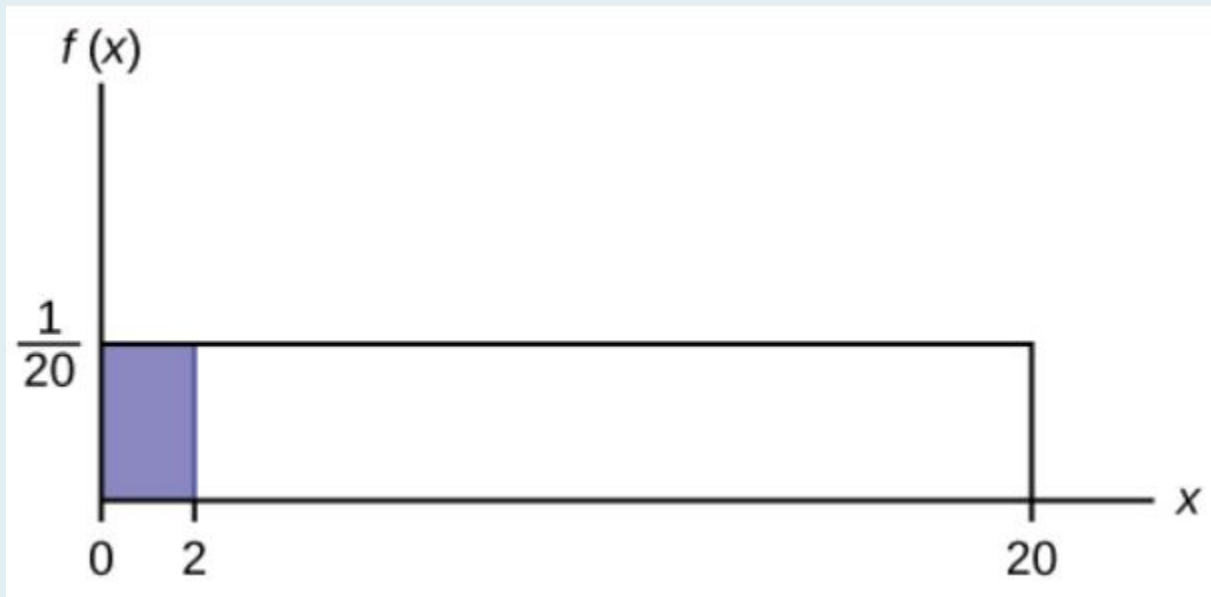
Suppose the graph below is the probability distribution of a continuous random variable. The graph is a horizontal line segment from $x = 0$ to $x = 20$. The height of the line is at $\frac{1}{20}$

for $0 \leq x \leq 20$. Note that the total area under the curve of $f(x)$, above the x -axis, from $x = 0$ to $x = 20$ is

$$\text{Area} = 20 \times \frac{1}{20} = 1$$



Suppose we want to find the area between the curve and the x -axis for $0 < x < 2$.



In this case, the area equals the area of a rectangle from $x = 0$ to $x = 2$. The area of a rectangle is $\text{base} \times \text{height}$, so

$$\text{Area} = (2 - 0) \times \frac{1}{20} = 0.1$$

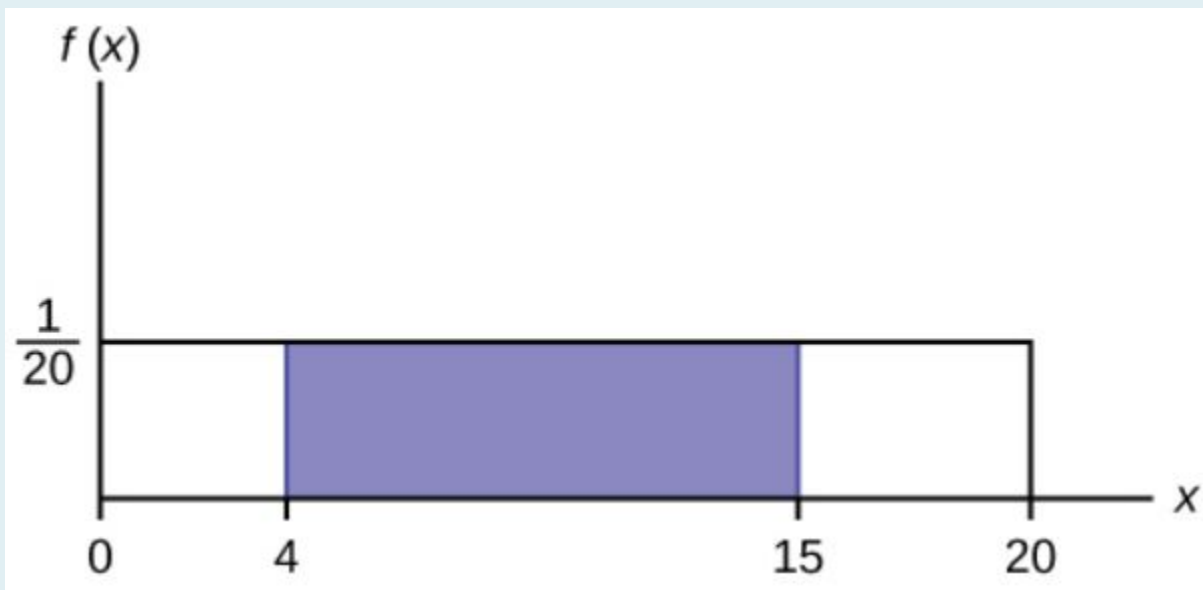
The area corresponds to the probability that the associated continuous random variable takes on a value between $x = 0$ and $x = 2$. Because the area is 0.1, the probability that $0 < x < 2$ is 0.1. Mathematically, we can write this as:

$$P(0 < x < 2) = 0.1$$

In other words, the probability that the random variable takes on a value between 0 and 2 is 0.1.

Suppose we want to find the probability that the random variable takes on a value between $x = 4$ and $x = 15$. This corresponds to the area under the curve in between $x = 4$ and $x = 15$.

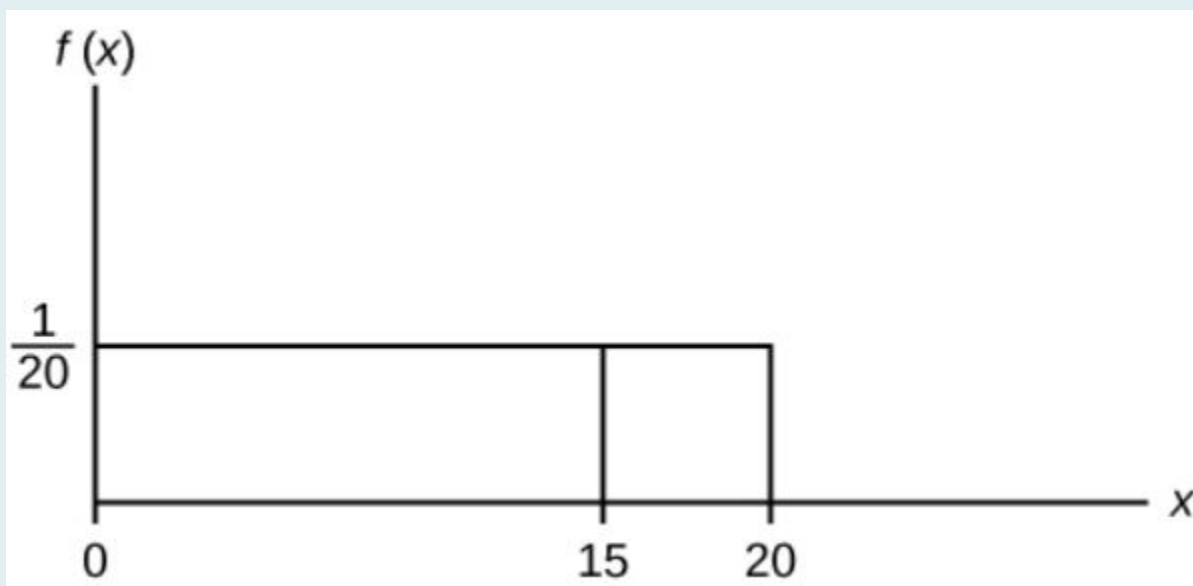
$$\text{Area} = (15 - 4) \times \frac{1}{20} = 0.55$$



So, the probability that the random variable takes on a value between 4 and 15 is 0.55.

Suppose we want to find $P(x = 15)$. This corresponds to the area above $x = 15$, which is just a vertical line. A vertical line has no width (or zero width). So

$$\text{Area} = 0 \times \frac{1}{20} = 0$$



So, the probability that the random variable equals 15 is 0.

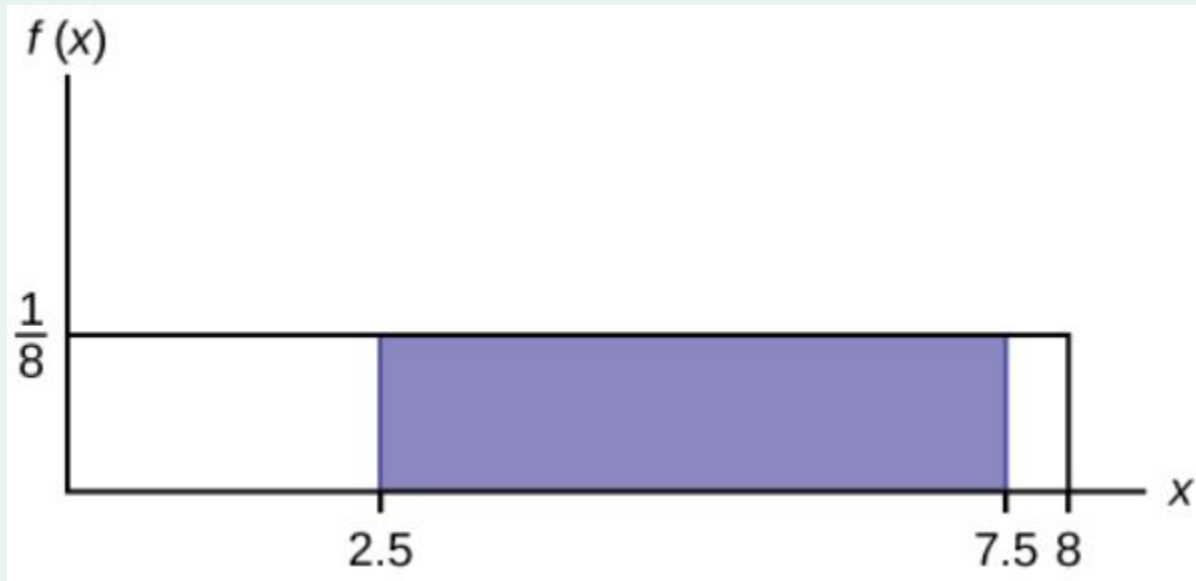
NOTE

The probability density function $f(x) = \frac{1}{20}$ used above is an example of a uniform distribution.

The graph of a uniform distribution is always a horizontal line.

TRY IT

Suppose the graph shown below is a probability distribution for a continuous random variable. Find the probability that the random variable takes on a value between 2.5 and 7.5.



Click to see Solution

$$\text{Area} = (7.5 - 2.5) \times \frac{1}{8} = 0.625$$



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=126#oembed-1>

Video: “Continuous probability distribution intro” by Khan Academy [9:58] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

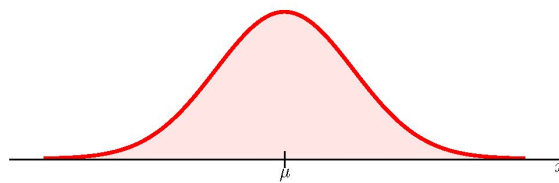
“5.2 Probability Distribution of a Continuous Random Variable” and “5.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

5.2 THE NORMAL DISTRIBUTION

LEARNING OBJECTIVES

- Describe properties of the normal distribution.
- Apply the Empirical Rule for normal distributions.

The normal distribution is the most important of all the distributions. It is widely used and even more widely abused. The normal distribution is a continuous probability distribution associated with a continuous random variable. The graph of a normal distribution is a symmetric, bell-shaped curve. This bell curve is used in almost all disciplines, including psychology, business, economics, the sciences, nursing, and, of course, mathematics. Most IQ scores are normally distributed. Often, real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.

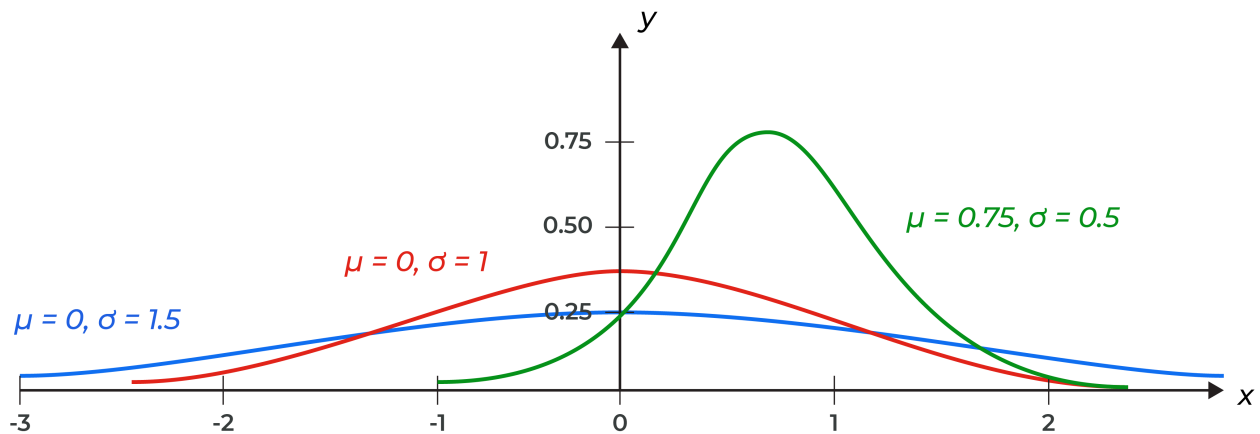


Properties of the normal distribution include:

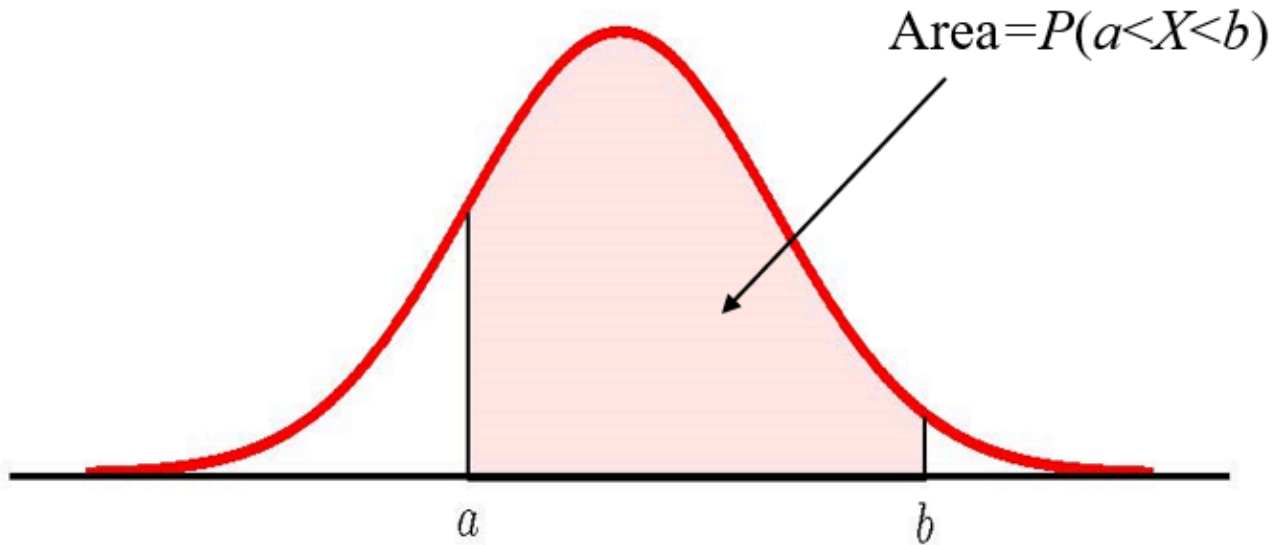
- The curve of a normal distribution is symmetric and bell-shaped.
- The center of a normal distribution is at the mean μ .
- In a normal distribution, the mean, the median, and the mode are equal.
- The curve is symmetric about a vertical line drawn through the mean.
- The tails of a normal distribution extend to infinity in both directions along the x -axis.

- The standard deviation, σ , of a normal distribution determines the shape, narrow or wide, of the curve.
- The total area under the curve of a normal distribution equals 1.

A normal distribution is completely determined by its mean μ and its standard deviation σ , which means there are an infinite number of normal distributions. The mean μ determines the center of the distribution—a change in the value of μ causes the graph to shift to the left or right. The standard deviation σ determines the shape, narrow or wide, of the bell. Because the area under the curve must equal one, a change in the standard deviation σ causes a change in the shape of the curve—the curve becomes fatter or skinnier depending on the value of σ .



As we saw in the previous section, the area under the curve of the normal distribution equals the probability that the corresponding normal random variable takes on a value within a given interval. That is, the probability that the normal random variable is in between a and b equals the area under the normal curve in between $x = a$ and $x = b$.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=132#oembed-1>

Video: “Continuous probability distribution intro” by Khan Academy [9:58] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

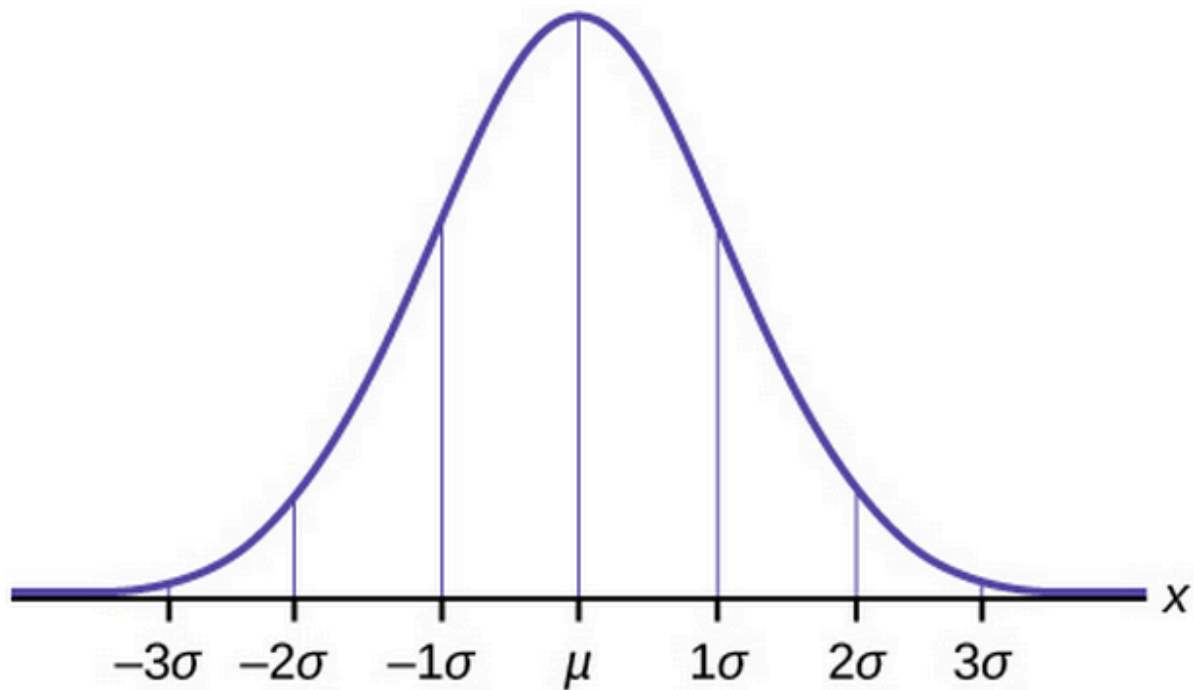
The Empirical Rule

For a normal distribution with mean μ and standard deviation σ , the **Empirical Rule** says the following:

- About 68\% of the values lie between $-1 \times \sigma$ and $+1 \times \sigma$ of the mean μ . In other words, about 68\% of the data fall with one standard deviation of the mean.
- About 95\% of the values lie between $-2 \times \sigma$ and $+2 \times \sigma$ of the mean μ . In other words, about 95\% of the data fall with two standard deviation of the mean.
- About 99.7\% of the values lie between $-3 \times \sigma$ and $+3 \times \sigma$ of the mean μ . In other words,

about 99.7\% of the data fall within three standard deviation of the mean

The empirical rule is also known as the **68 – 95 – 99.7 rule**.



EXAMPLE

Suppose a normal distribution has a mean of 50 and a standard deviation of 6.

The Empirical Rule says:

- About 68\% of the values lie between $-1 \times \sigma = -1 \times 6 = -6$ and $1 \times \sigma = 1 \times 6 = 6$ of the mean 50. The values $50 - 6 = 44$ and $50 + 6 = 56$ are within one standard deviation of the mean 50. So 68\% of the values in this distribution are between 44 and 56.
- About 95\% of the values lie between $-2 \times \sigma = -2 \times 6 = -12$ and $2 \times \sigma = 2 \times 6 = 12$ of the mean 50. The values $50 - 12 = 38$ and $50 + 12 = 62$ are

within two standard deviations of the mean 50. So 95% of the values in the distribution are between 38 and 62.

- About 99.7% of the values lie between $-3 \times \sigma = -3 \times 6 = -18$ and $3 \times \sigma = 3 \times 6 = 18$ of the mean 50. The values $50 - 18 = 32$ and $50 + 18 = 68$ are within three standard deviations of the mean 50. So 99.7% of the values in the distribution are between 32 and 68.

TRY IT

Suppose a normal distribution has a mean of 25 and a standard deviation of 5. 68% of the values lie between what two numbers?

Click to see Solution

Between $25 + (-1) \times 5 = 20$ and $25 + 1 \times 5 = 30$.

EXAMPLE

From 1984 to 1985, the height of 15 to 18-year-old males from Chile follows a normal distribution with mean 172.36cm and standard deviation 6.34cm.

1. About 68% of the heights of 15 to 18-year-old males in Chile from 1984 to 1985 lie between

what two values?

2. About 95\% of the heights of 15 to 18-year-old males in Chile from 1984 to 1985 lie between what two values?
3. About 99.7\% of the heights of 15 to 18-year-old males in Chile from 1984 to 1985 lie between what two values?

Solution

1. $\mu + (-1) \times \sigma = 172.36 + (-1) \times 6.34 = 166.02$ and $\mu + 1 \times \sigma = 172.36 + 1 \times 6.34 = 178.70$
2. $\mu + (-2) \times \sigma = 172.36 + (-2) \times 6.34 = 159.68$ and $\mu + 2 \times \sigma = 172.36 + 2 \times 6.34 = 185.04$
3. $\mu + (-3) \times \sigma = 172.36 + (-1) \times 6.34 = 153.34$ and $\mu + 3 \times \sigma = 172.36 + 2 \times 6.34 = 191.36$

TRY IT

The scores on a college entrance exam have an approximate normal distribution with a mean of 52 points and a standard deviation of 11 points.

1. About 68\% of the exam scores lie between what two values?
2. About 95\% of the exam scores lie between what two values?
3. About 99.7\% of the exam scores lie between what two values?

Click to see Solution

1. About 68\% of the scores lie between the values $52 + (-1) \times 11 = 41$ and $52 + 1 \times 11 = 63$.
2. About 95\% of the values lie between the values $52 + (-2) \times 11 = 30$ and $52 + 2 \times 11 = 74$.
3. About 99.7\% of the values lie between the values $52 + (-3) \times 11 = 19$ and

$$52 + 3 \times 11 = 85.$$



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=132#oembed-2>

Video: “ck12.org normal distribution problems: Empirical rule | Probability and Statistics | Khan Academy” by Khan Academy [10:25] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

Click to see Answer

61

2. About what percent of the values of a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

Click to see Answer

68\%

3. About what percent of the values of a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

Click to see Answer

95\%

4. About what percent of the values of a normal distribution lie between the second and third standard deviations (both sides)?

Click to see Answer

4.7\%

5. Suppose X is a normal random variable with mean 15 and standard deviation 3. Between what two values does 68\% of the data lie?

Click to see Answer

15 and 18

6. Suppose X is a normal random variable with mean -3 and standard deviation 1.
- Between what two values does 95\% of the data lie?
 - Between what two values does 34\% of the data lie?

Click to see Answer

- -5 and -1
- -4 and -3 or -3 and -2

7. About what percent of the values of a normal distribution lie between the mean and three standard deviations?

Click to see Answer

49.85\%

8. About what percent of the values of a normal distribution lie between the mean and one standard deviation?

Click to see Answer

34\%

9. About what percent of the values of a normal distribution lie between the first and second standard deviations from the mean (both sides)?

Click to see Answer

27\%

10. About what percent of the values of a normal distribution lie between the first and third standard deviations (both sides)?

Click to see Answer

31.7\%

11. The salaries of employees of a large technology company follow a normal distribution with a mean of \$65, 000 and a standard deviation of \$1, 700.
- 68\% of the salaries lie between what two values?
 - 99.7\% of the salaries lie between what two values?
 - 2.5\% of the salaries lie above what value?
 - 16\% of the salaries lie below what value?
 - 97\% of the salaries lie above what value?
 - What percentage of salaries lie between \$61, 600 and \$68, 400?
 - What percentage of salaries lie above \$70, 100?
 - What percentage of salaries lie below \$61, 600?
 - What percentage of salaries lie between \$65, 000 and \$68, 400?
 - What percentage of salaries lie above \$66, 700?

Click to see Answer

- \$63, 300 and \$66, 700
- \$59, 900 and \$70, 100
- \$68, 400
- \$63, 300
- \$61, 600
- 95\%
- 0.15\%
- 2.5\%
- 47.5\%
- 16\%

12. A factory produces light bulbs. The lifespan of the light bulbs follows a normal distribution with a mean of 1200 hours and a standard deviation of 100 hours.
- 95\% of the bulbs have lifespans between what two values?
 - 16\% of the bulbs have lifespans above what value?
 - 0.15\% of the bulbs have lifespans below what value?
 - 99.85\% of the bulbs have lifespans below what value?
 - 84\% of the bulbs have lifespans above what value?
 - What percentage of bulbs have lifespans between 1100 hours and 1300 hours?
 - What percentage of bulbs have lifespans below 1000 hours?
 - What percentage of bulbs have lifespans above 1500 hours?
 - What percentage of bulbs have lifespans below 1400 hours?
 - What percentage of bulbs have lifespans between 900 hours and 1200 hours?
 - What percentage of bulbs have lifespans between 1200 hours and 1300 hours?

Click to see Answer

- 1000 hours and 1400 hours
 - 1300 hours
 - 900 hours
 - 1500 hours
 - 1100 hours
 - 68\%
 - 2.5\%
 - 0.15\%
 - 97.55\%
 - 49.85\%
 - 34\%
13. The distribution of the heights of adult women follows a normal distribution with a mean of 163 cm and a standard deviation of 8 cm.
- 99.7\% of women have heights between what two values?
 - 84\% of women have heights above what value?
 - 0.15\% of women have heights above what value?
 - 97.5\% of women have heights below what value?
 - What percentage of women have heights between 155 cm and 171 cm?
 - What percentage of women have heights below 139 cm?
 - What percentage of women have heights above 171 cm?

- h. What percentage of women have heights above 147 cm?
- i. What percentage of women have heights between 147 cm and 163 cm?

Click to see Answer

- a. 139 cm and 187 cm
- b. 155 cm
- c. 187 cm
- d. 179 cm
- e. 68\%
- f. 0.15\%
- g. 16\%
- h. 97.5\%
- i. 47.5\%

“5.3 The Normal Distribution” and “5.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

5.3 THE STANDARD NORMAL DISTRIBUTION

LEARNING OBJECTIVES

- Recognize the standard normal probability distribution and apply it appropriately.

The **standard normal distribution** is the normal distribution with $\mu = 0$ and $\sigma = 1$. The normal random variable associated with the standard normal distribution is denoted Z .

For any normal distribution with mean μ and standard deviation σ , a **z -score** is the number of the standard deviations a value x is from the mean. For example, if a normal distribution has $\mu = 5$ and $\sigma = 2$, then for $x = 11$

$$11 = x = \mu + z \times \sigma = 5 + 3 \times 2$$

In this case, $z = 3$. We would say that 11 is three standard deviations above (or to the right of) the mean.

The standard normal distribution is the normal distribution of these **standardized z -scores**. For any normal distribution with mean μ and standard deviation σ , we can transform the normal distribution to the standard normal distribution using the formula

$$z = \frac{x - \mu}{\sigma}$$

where x is a value from the normal distribution. The z -score is the number of standard deviations the value x is above (to the right of) or below (to the left of) the mean μ . Values of x that are larger than the mean have positive z -scores, and values of x that are smaller than the mean have negative z -scores. If x equals the mean, then x has a z -score of zero.

EXAMPLE

Suppose a normal distribution has mean $\mu = 5$ and standard deviation $\sigma = 6$.

For $x = 17$, the z -score is

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{17 - 5}{6} \\ &= 2 \end{aligned}$$

This says that $x = 17$ is **two standard deviations** ($2 \times \sigma$) above or to the right of the mean $\mu = 5$. Notice that $x = 5 + 2 \times 6 = 17$

For $x = 1$, the z -score is

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{1 - 5}{6} \\ &= -0.666\dots \end{aligned}$$

This says that $x = 1$ is **0.666... standard deviations** ($-0.666\dots \times \sigma$) below or to the left of the mean $\mu = 5$. Notice that $x = 5 + (-0.666\dots) \times 6 = 1$

NOTES

- When z is positive, x is above or to the right of the mean μ . In other words, x is greater than μ .
- When z is negative, x is below or to the left of the mean μ . In other words, x is less than μ .

TRY IT

What is the z -score of $x = 1$ for a normal distribution with $\mu = 12$ and $\sigma = 3$?

Click to see Solution

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{1 - 12}{3} \\ &= -3.666\ldots \end{aligned}$$



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=134#oembed-1>

Video: “ck12.org normal distribution problems: z-score | Probability and Statistics | Khan Academy” by Khan Academy [7:48] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

EXAMPLE

Some doctors believe that a person can lose five pounds, on average, in a month by reducing his or her fat intake and by exercising consistently. Suppose the amount of weight (in pounds) a person loses in a month has a normal distribution with $\mu = 5$ and $\sigma = 2$. Fill in the blanks.

1. Suppose a person lost ten pounds in a month. The z -score when $x = 10$ pounds is $z = 2.5$ (verify). This z -score tells us that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean _____. (What is the mean?).
2. Suppose a person gained three pounds (a negative weight loss). Then $z =$ _____. This z -score tells us that $x = -3$ is _____ standard deviations to the _____ (right or left) of the mean.

Solution

1. This z -score tells us that $x = 10$ is 2.5 standard deviations to the **right** of the mean 5.
2. $z = -4$. This z -score tells us that $x = -3$ is 4 standard deviations to the **left** of the mean.

EXAMPLE

Suppose X is a normal random variable with $\mu = 5$ and $\sigma = 6$ and Y is a normal random variable with $\mu = 2$ and $\sigma = 1$.

Suppose $x = 17$:

$$\begin{aligned}
 z &= \frac{x - \mu}{\sigma} \\
 &= \frac{17 - 5}{6} \\
 &= 2
 \end{aligned}$$

The z -score for $x = 17$ is $z = 2$, which means that 17 is 2 standard deviations to the right of the mean $\mu = 5$.

Suppose $y = 4$:

$$\begin{aligned}
 z &= \frac{y - \mu}{\sigma} \\
 &= \frac{4 - 2}{1} \\
 &= 2
 \end{aligned}$$

The z -score for $y = 4$ is $z = 2$, which means that 4 is 2 standard deviations to the right of the mean $\mu = 2$.

Therefore, $x = 17$ and $y = 4$ are both two (of **their own**) standard deviations to the right of their respective means. In other words, compared to the mean of their corresponding distributions, $x = 17$ and $y = 4$ have the same **relative** position.

NOTE

The z -score allows us to compare data that are scaled differently by considering the data's position relative to its mean. To understand the concept, suppose X represents weight gains for one group of people who are trying to gain weight in a six-week period, and Y measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Because $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same standardized weight gain relative to their means.

TRY IT

Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of 4 points. Suppose Jerome scores 10 points in a game. The z -score when $x = 10$ is -1.5 . This score tells us that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

Click to see Solution

1.5, left, 16

EXAMPLE

The heights of 15 to 18-year-old males from Chile from 2009 to 2010 follows a normal distribution with mean 170 cm and standard deviation 6.28 cm.

1. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The z -score when $x = 168$ cm is $z =$ _____. This z -score tells us that $x = 168$ is _____ standard deviations to the _____ (right or left) of the mean _____. (What is the mean?).
2. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z -score of $z = 1.27$. What is the male's height? The z -score ($z = 1.27$) tells us that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Solution

1. -0.32 , 0.32 , left, 170

2. 177.98, 1.27, right

TRY IT

The heights of 15 to 18-year-old males from Chile from 2009 to 2010 follows a normal distribution with mean 170 cm and standard deviation 6.28 cm.

1. Suppose a 15 to 18-year-old male from Chile was 176 cm tall from 2009 to 2010. The z -score when $x = 176$ cm is $z =$ _____. This z -score tells us that $x = 176$ cm is _____ standard deviations to the _____ (right or left) of the mean _____. (What is the mean?).
2. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z -score of $z = -2$. What is the male's height? The z -score ($z = -2$) tells us that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Click to see Solution

1. $z = \frac{x - \mu}{\sigma} = \frac{176 - 170}{6.28} = 0.96$. This z -score tells us that $x = 176$ cm is 0.96 standard deviations to the right of the mean 170 cm.
2. $x = \mu + z \times \sigma = 170 + (-2) \times 6.28 = 157.44$ cm. The z -score ($z = -2$) tells us that the male's height is 2 standard deviations to the left of the mean.

EXAMPLE

From 2009 to 2010, the heights of 15 to 18-year-old males from Chile from 2009 to 2010 follows a normal distribution with a mean of 170 cm and a standard deviation of 6.28 cm. Let X be the height of a 15 to 18-year-old male from Chile in 2009 to 2010.

From 1984 to 1985, the heights of 15 to 18-year-old males from Chile follows a normal distribution with mean 172.36 cm and standard deviation 6.34 cm. Let Y be the height of a 15 to 18-year-old male from Chile in 1984 to 1985.

Find the z -scores for $x = 160.58$ cm and $y = 162.85$ cm. Interpret each z -score. What can you say about $x = 160.58$ cm and $y = 162.85$ cm?

Solution

The z -score for $x = 160.58$ is

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{160.58 - 170}{6.28} \\ &= -1.5 \end{aligned}$$

The z -score for $y = 162.85$ is

$$\begin{aligned} z &= \frac{y - \mu}{\sigma} \\ &= \frac{162.85 - 172.36}{6.34} \\ &= -1.5 \end{aligned}$$

Both $x = 160.58$ and $y = 162.85$ deviate the same number of standard deviations from their respective means and in the same direction.

TRY IT

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT followed a normal distribution with a mean of 496 and a standard deviation of 114.

Find the z -scores for Student 1 with a score of 325 and for Student 2 with a score of 366.21. Interpret each z -score. What can we say about these two students' scores?

Click to see Solution

For Student 1:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{325 - 496}{114} \\ &= -1.5 \end{aligned}$$

For Student 2:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{366.21 - 496}{114} \\ &= -1.138\dots \end{aligned}$$

Student 2 scored closer to the mean than Student 1, and because they both had negative z -scores, Student 2 had the better score.

Exercises

1. What does a z -score measure?

Click to see Answer

The number of standard deviations a value is from the mean of the distribution.

2. What is the z -score of $x = 12$ if it is two standard deviations to the right of the mean?

Click to see Answer

2

3. What is the z -score of $x = 9$ if it is 1.5 standard deviations to the left of the mean?

Click to see Answer

−1.5

4. What is the z -score of $x = -2$ if it is 2.78 standard deviations to the right of the mean?

Click to see Answer

−2.78

5. What is the z -score of $x = 7$ if it is 0.133 standard deviations to the left of the mean?

Click to see Answer

−0.133

6. Suppose X is a normal random variable with a mean of 2 and a standard deviation of 6. What value of x has a z -score of 3?

Click to see Answer

20

7. Suppose X is a normal random variable with a mean of 8 and a standard deviation of 1. What value of x has a z -score of −2.25?

Click to see Answer

5.75

8. Suppose X is a normal random variable with a mean of 9 and a standard deviation of 5. What value of x has a z -score of -0.5 ?

Click to see Answer

6.5

9. Suppose X is a normal random variable with a mean of 2 and a standard deviation of 3. What value of x has a z -score of -0.67 ?

Click to see Answer

-0.01

10. Suppose X is a normal random variable with a mean of 4 and a standard deviation of 2. What value of x is 1.5 standard deviations to the left of the mean?

Click to see Answer

1

11. Suppose X is a normal random variable with a mean of 4 and a standard deviation of 2. What value of x is 2 standard deviations to the right of the mean?

Click to see Answer

8

12. Suppose X is a normal random variable with a mean of 8 and a standard deviation of 9. What value of x is 0.67 standard deviations to the left of the mean?

Click to see Answer

1.97

13. Suppose X is a normal random variable with a mean of -1 and a standard deviation of 2. What is the z -score of $x = 2$?

Click to see Answer

1.5

14. Suppose X is a normal random variable with a mean of 12 and a standard deviation of 6.

What is the z -score of $x = 3$?

Click to see Answer

−1.5

15. Suppose X is a normal random variable with a mean of 9 and a standard deviation of 3. What is the z -score of $x = 9$?

Click to see Answer

0

16. Suppose a normal distribution has a mean of 6 and a standard deviation of 1.5. What is the z -score of $x = 3.975$?

Click to see Answer

−1.35

17. In a normal distribution, $x = 5$ and $z = -1.25$. This tells you that $x = 5$ is ____ standard deviations to the ____ (right or left) of the mean.

Click to see Answer

1.25, left

18. In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is ____ standard deviations to the ____ (right or left) of the mean.

Click to see Answer

0.67, right

19. In a normal distribution, $x = -2$ and $z = 6$. This tells you that $x = -2$ is ____ standard deviations to the ____ (right or left) of the mean.

Click to see Answer

6, right

20. In a normal distribution, $x = -5$ and $z = -3.14$. This tells you that $x = -5$ is ____ standard deviations to the ____ (right or left) of the mean.

Click to see Answer

3.14, left

21. In a normal distribution, $x = 6$ and $z = -1.7$. This tells you that $x = 6$ is ____ standard deviations to the ____ (right or left) of the mean.

Click to see Answer

1.7, left

22. The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.
- What is the median recovery time?
 - What is the z -score for a patient who takes ten days to recover?
 - What is the z -score for a patient who takes three days to recover?

Click to see Answer

- 5.3 days
- 2.238
- 1.095

23. The heights of the 430 National Basketball Association players were listed on team rosters at the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with a mean of 79 inches and a standard deviation of 3.89 inches.
- Calculate the z -score for a player with a height of 77 inches. Interpret the z -score.
 - Calculate the z -score for a player with a height of 85 inches. Interpret the z -score.
 - If an NBA player reported his height had a z -score of 3.5, would you believe him? Explain your answer.

Click to see Answer

- 0.514; The height of 77 inches is 0.514 standard deviations to the left of the mean.

- b. 1.542; The height of 85 inches is 1.542 standard deviations to the right of the mean.
- c. No, because the player would need to be 92.615 inches (over 7 feet, 7 inches) tall.

24. The systolic blood pressure (given in millimetres) of males has an approximately normal distribution with a mean of 125 and a standard deviation of 14. Systolic blood pressure for males follows a normal distribution.
- a. Calculate the z -scores for the male systolic blood pressures 100 and 150 millimetres.
 - b. If a male friend of yours said he thought his systolic blood pressure was 2.5 standard deviations below the mean but that he believed his blood pressure was between 100 and 150 millimetres, what would you say to him?

Click to see Answer

- a. $-1.786, 1.786$
 - b. His systolic blood pressure cannot be between 100 and 150 millimetres because at 2.5 standard deviations below the mean, his blood pressure is 90 millimetres.
25. Kyle's doctor told him that the z -score for his systolic blood pressure is 1.75. The systolic blood pressure (given in millimetres) of males has an approximately normal distribution with a mean of 125 and a standard deviation of 14.
- a. Which of the following is the best interpretation of this standardized score?
 - i. Kyle's systolic blood pressure is 175.
 - ii. Kyle's systolic blood pressure is 1.75 times the average blood pressure of men his age.
 - iii. Kyle's systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.
 - iv. Kyle's systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.
 - b. Calculate Kyle's blood pressure.

Click to see Answer

- a. iv
 - b. 149.5
26. Height and weight are two measurements used to track a child's development. The World Health Organization measures child development by comparing the weights of children who are the same height and the same gender. In 2009, weights for all 80 cm girls in the reference population had a mean of 10.2 kg and a standard deviation of 0.8 kg. Weights are normally

distributed. Calculate the z -scores that correspond to the following weights and interpret them.

- a. 11 kg
- b. 7.9 kg
- c. 12.2 kg

Click to see Answer

- a. 1; The weight of 11 kg is 1 standard deviations to the right of the mean.
- b. -2.875 ; The weight of 7.9 kg is 2.875 standard deviations to the left of the mean.
- c. 2.5; The weight of 12.2 kg is 2.5 standard deviations to the right of the mean.

27. In 2005, 1,475,623 students heading to college took the SAT. The distribution of scores in the math section of the SAT follows a normal distribution with a mean of 520 and a standard deviation of 115.
- a. Calculate the z -score for an SAT score of 720. Interpret it using a complete sentence.
 - b. What math SAT score is 1.5 standard deviations above the mean? What can you say about this SAT score?
 - c. For 2012, the SAT math test had a mean of 514 and a standard deviation of 117. The ACT math test is an alternative to the SAT and is approximately normally distributed with a mean of 21 and a standard deviation of 5.3. If one person took the SAT math test and scored 700 and a second person took the ACT math test and scored 30, who did better with respect to the test they took?

Click to see Answer

- a. 1.739; The score of 720 is 1.739 standard deviations to the right of the mean.
- b. 692.5; The score of 692.5 kg is 1.5 standard deviations to the right of the mean.
- c. SAT z -score is 1.565; ACT z -score is 1.698; The person who wrote the ACT test did better because the z -score for their score is higher.

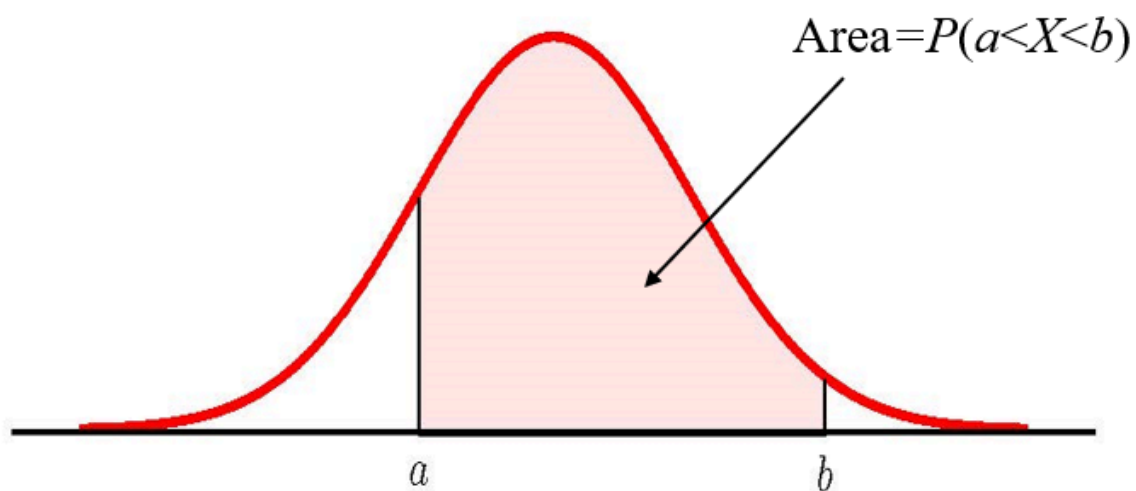
“5.4 The Standard Normal Distribution” and “5.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

5.4 CALCULATING PROBABILITIES FOR A NORMAL DISTRIBUTION

LEARNING OBJECTIVES

- Recognize the normal probability distribution and apply it appropriately.
- Calculate probabilities associated with a normal distribution.

Probabilities for a normal random variable X equal the area under the corresponding normal distribution curve. The probability that the value for X falls in between the values $x = a$ and $x = b$ is the area under the normal distribution curve to the right of $x = a$ and to the left of $x = b$.



CALCULATING NORMAL PROBABILITIES IN EXCEL

To calculate probabilities associated with normal random variables in Excel, use the **norm.dist(x,μ,σ,logic operator)** function.

- For **x**, enter the value for x .
- For **μ**, enter the mean of the normal distribution.
- For **σ**, enter the standard deviation of the normal distribution.
- For the logic operator, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

The output from the **norm.dist** function is the probability that $X < x$. That is the output from the **norm.dist** function is the area to the **left** of the value of **x**.

Visit the Microsoft page for more information about the **norm.dist** function.

NOTE

The **norm.dist** function always tells us the area to the left of the value entered for **x**.

- To find the area to the right of the value of **x**, we use **1-norm.dist(x,μ,σ,true)**. This corresponds to the probability that $X > x$.
- To find the area in between **x₁** and **x₂** with $x_1 < x_2$, we use **norm.dist(x₂,μ,σ,true)-norm.dist(x₁,μ,σ,true)**. This corresponds to the probability that $x_1 < X < x_2$.

An alternative approach in Excel is to use the **norm.s.dist(z,true)** function. In the **norm.s.dist** function, we enter the **z**-score for the corresponding value of **x** and the output will be the area to the left **x**.

CALCULATING x -VALUES FOR A NORMAL DISTRIBUTION IN EXCEL

Given the area to the left of an (unknown) x -value, use the **norm.inv(probability, μ , σ)** function.

- For **probability**, enter the area to the **left** of x .
- For μ , enter the mean of the normal distribution.
- For σ , enter the standard deviation of the normal distribution.

The output from the **norm.inv** function is the value of x so that the area to left of x equals the given probability. That is the output from the **norm.inv** function is the value of x so that the $P(X < x) = \text{probability}$.

Visit the Microsoft page for more information about the **norm.inv** function.

NOTE

The **norm.inv** function requires that we enter the area to the **left** of the unknown x -value. If we are given the area to the **right** of the unknown x -value, we enter **1-area to the right** for the probability in the **norm.inv** function. That is, given the area to the **right** of the x -value, we use **norm.inv(1-area, μ , σ)**.

EXAMPLE

The final exam scores in a statistics class are normally distributed with a mean of **63** and a standard deviation of **5**.

1. Find the probability that a randomly selected student scored more than **65** on the exam.
2. Find the probability that a randomly selected student scored less than **75** on the exam.
3. 90\% of the students scored less than what value?
4. 30\% of the students scored more than what value?

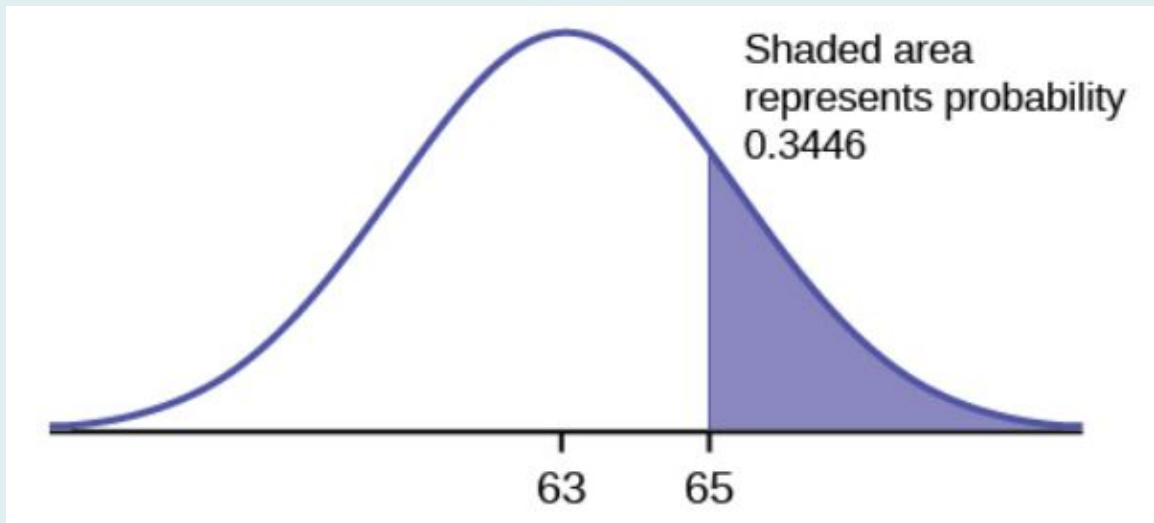
Solution

Let X be the scores on the final exam.

1. We want to find $P(X > 65)$:

Function	1-norm.dist
Field 1	65
Field 2	63
Field 3	5
Field 4	true
Answer	0.3446

The probability that a student scores more than **65** is **0.3446** (or 34.46\%).



2. We want to find $P(X < 75)$:

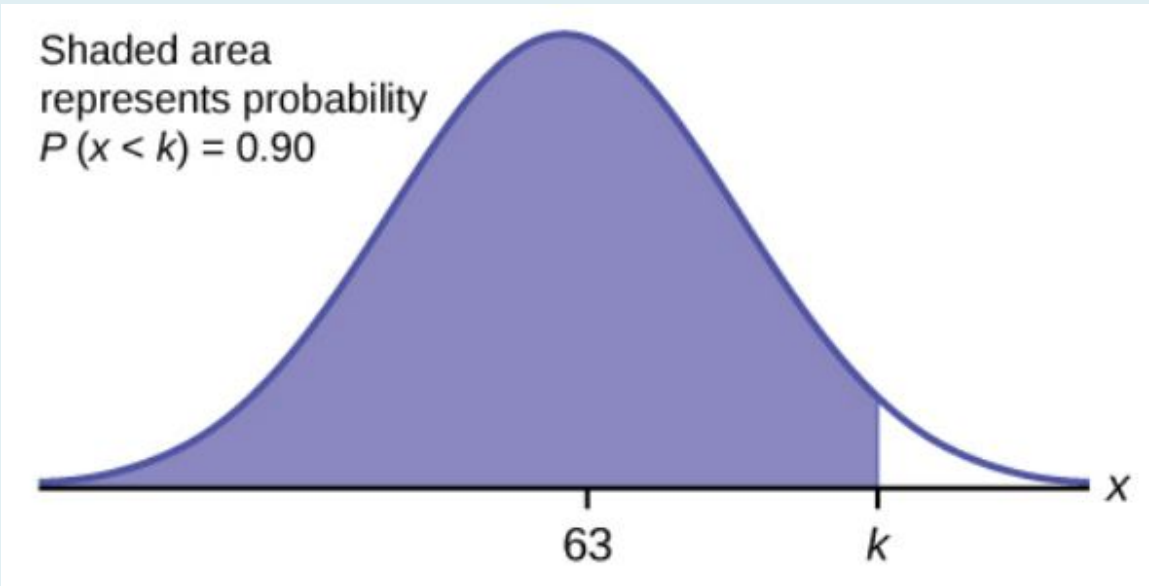
Function	norm.dist
Field 1	75
Field 2	63
Field 3	5
Field 4	true
Answer	0.9918

The probability that a student scores less than 75 is 0.9918 (or 99.18%).

3. We want to find the value of x so that the area to the left of x is 0.9.

Function	norm.inv
Field 1	0.9
Field 2	63
Field 3	5
Answer	69.41

90% of the students scored below 69.41 points on the exam.



4. We want to find the value of x so that the area to the right of x is 0.3. This is the same as finding the value of x so that the area to the left of x is 0.7 ($1 - 0.3$).

Function	norm.inv
Field 1	0.7
Field 2	63
Field 3	5
Answer	65.62

30\% of the students scored more 65.62 points on the exam.

TRY IT

The golf scores for a school team are normally distributed with a mean of 68 and a standard deviation of 3.

1. Find the probability that a randomly selected golfer scored less than **65**.
2. Find the probability that a randomly selected golfer scored more than **72**.

Click to see Solution

1.

Function	norm.dist
Field 1	65
Field 2	68
Field 3	3
Field 4	true
Answer	0.1587

2.

Function	1-norm.dist
Field 1	72
Field 2	68
Field 3	3
Field 4	true
Answer	0.0912

EXAMPLE

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is **2** hours per day. Assume the times for entertainment are normally distributed, and the standard deviation for the times is **0.5** hour.

1. Find the probability that a household personal computer is used for entertainment between

- 1.8 and 2.75 hours per day.
2. Find the number of hours per day for the bottom 25\% of households using a personal computer for entertainment.

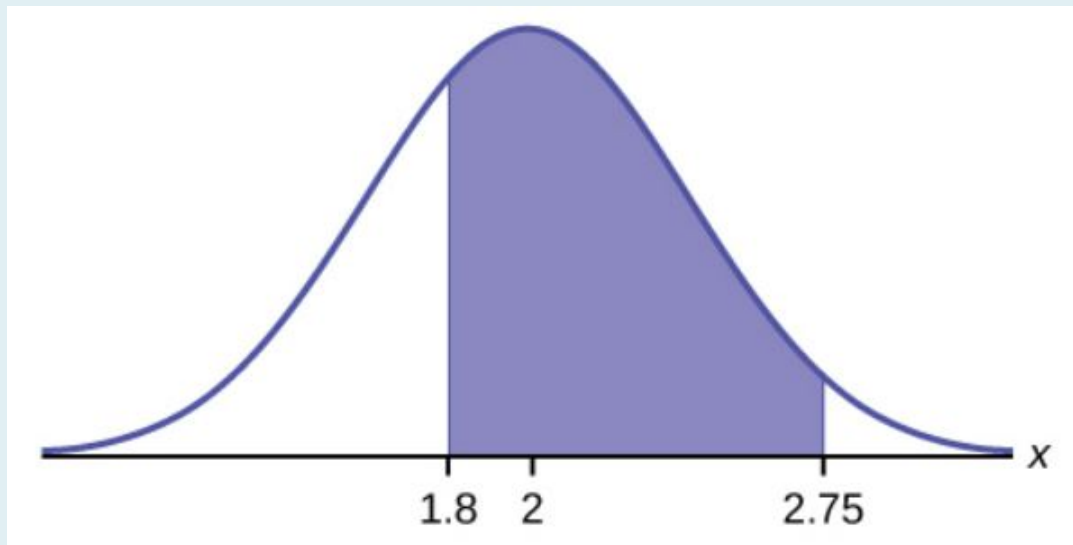
Solution

Let X be the amount of time (in hours) a household personal computer is used for entertainment.

1. We want to find $P(1.8 < X < 2.75)$.

Function	norm.dist	-norm.dist
Field 1	2.75	1.8
Field 2	2	2
Field 3	0.5	0.5
Field 4	true	true
Answer	0.5886	

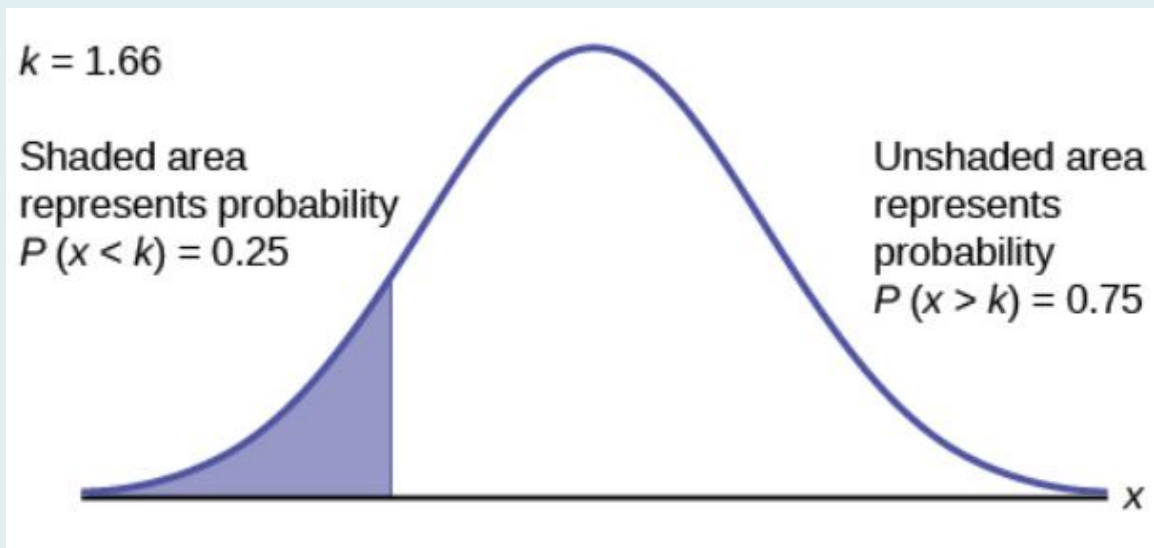
The probability a household computer is used for entertainment between 1.8 and 2.75 hours a day is 0.5886 (or 58.86\%).



2. We need to find the value x so that 25\% of the number of hours as less than this value.

Function	norm.inv
Field 1	0.25
Field 2	2
Field 3	0.5
Answer	1.66

25\% of the value are less than **1.66** hours.



TRY IT

The golf scores for a school team are normally distributed with a mean of **68** and a standard deviation of **3**. Find the probability that a golfer scored between **66** and **70**.

Click to see Solution

Function	norm.dist	-norm.dist
Field 1	70	66
Field 2	68	68
Field 3	3	3
Field 4	true	true
Answer	0.4950	

EXAMPLE

There are approximately one billion smartphone users in the world today. In the United States, the ages of smartphone users from 13 to 55+ follow a normal distribution with approximate mean and standard deviation of **36.9** years and **13.9** years, respectively.

1. Determine the probability that a random smartphone user in the age range 13 to 55+ is between **23** and **64.7** years old.
2. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most **50.8** years old.
3. 80\% of the users in the age range 13 to 55+ are less than what age?
4. 40\% of the ages that range from 13 to 55+ are at least what age?

Solution

1.

Function	norm.dist	-norm.dist
Field 1	64.7	23
Field 2	36.9	36.9
Field 3	13.9	13.9
Field 4	true	true
Answer	0.8186	

The probability a smartphone user is between **23** and **64.7** years of age is **0.8186** (or 81.86\%).

2.

Function	norm.dist
Field 1	50.8
Field 2	36.9
Field 3	13.9
Field 4	true
Answer	0.8413

The probability that a smartphone user is less than **50.8** years of age is **0.8413** (or 84.13\%).

3.

Function	norm.inv
Field 1	0.8
Field 2	36.9
Field 3	13.9
Answer	48.6

80\% of the smartphone users in the age range 13 – 55+ are **48.6** years old or less.

4.

Function	norm.inv
Field 1	0.6
Field 2	36.9
Field 3	13.9
Answer	40.42

40\% of the smartphone users in the age range 13 – 55+ are older than 40.42 years of age.

TRY IT

There are approximately one billion smartphone users in the world today. In the United States, the ages of smartphone users from 13 to 55+ follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

- 1. 30\% of smartphone users are older than what age?
- 2. What is the probability that the age of a randomly selected smartphone user in the range of 13 to 55+ is less than 27 years old?

Click to see Solution

1.

Function	norm.inv
Field 1	0.7
Field 2	36.9
Field 3	13.9
Answer	44.19

2.

Function	norm.dist
Field 1	27
Field 2	36.9
Field 3	13.9
Field 4	true
Answer	0.2382

EXAMPLE

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

1. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm.
2. 90\% of the diameters of the mandarin oranges are less than what value?
3. 35\% of the diameters of the mandarin oranges are greater than what value?

Solution

1.

Function	1-norm.dist
Field 1	6
Field 2	5.85
Field 3	0.24
Field 4	true
Answer	0.2660

The probability an orange has a diameter greater than 6 cm is 0.2660 (or 26.60\%).

2.

Function	norm.inv
Field 1	0.9
Field 2	5.85
Field 3	0.24
Answer	6.16

90\% of the diameters of the oranges are less than 6.16 cm.

3.

Function	norm.inv
Field 1	0.65
Field 2	5.85
Field 3	0.24
Answer	5.94

35\% of the diameters of the oranges are greater than 5.94 cm.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=140#oembed-1>

Video: “Excel 2013 Statistical Analysis #39: Probabilities for Normal (Bell) Probability Distribution” by excelisfun [24:08] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. If the area to the left of x in a normal distribution is 0.123, what is the area to the right of x ?

Click to see Answer

0.877

2. If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x ?

Click to see Answer

0.457

3. Suppose X is a normal random variable with a mean of 54 and a standard deviation of 8.
- Find the probability that $x > 56$.
 - Find the probability that $x < 30$.
 - Find the probability that $40 < x < 50$.
 - 80\% of the x -values are less than what value?
 - 40\% of the x -values are greater than what value?

Click to see Answer

- 0.4013
 - 0.0014
 - 0.2685
 - 60.73
 - 56.03
4. The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years.
- Find the probability that a CD player will last between 2.8 and 6 years.
 - Find the probability that a CD player will last more than 7.5 years.
 - A CD player is guaranteed for three years. Find the probability that a CD player will break down during the guarantee period.
 - 70\% of the CD players last how long?
 - 10\% of the CD players will last more than how many years?

Click to see Answer

- 0.7694
 - 0.0045
 - 0.1987
 - 4.78 years
 - 5.77 years
5. The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.
- What is the probability of spending more than two days in recovery?
 - What is the probability of spending between four and seven days in recovery?
 - What percentage of patients spend fewer than three days in recovery?
 - 10\% of the recovery times are larger than what value?

Click to see Answer

- a. 0.9420
- b. 0.5230
- c. 0.1367
- d. 7.99 days

6. The length of time it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of 5 minutes and a standard deviation of 2 minutes.
- a. Would you be surprised if it took less than one minute to find a parking space? Justify your answer
 - b. Find the probability that it takes at least eight minutes to find a parking space.
 - c. 70\% of the time, it takes more than how many minutes to find a parking space?
 - d. 25\% of the time, it takes fewer than how many minutes to find a parking space?

Click to see Answer

- a. Yes, because the probability of finding a parking space in less than one minute is only 0.0228
- b. 0.0668
- c. 3.95 minutes
- d. 3.65 minutes

7. According to a study done by De Anza students, the height for Asian adult males is normally distributed with an average of 66 inches and a standard deviation of 2.5 inches.
- a. Find the probability that a randomly selected Asian adult male is between 65 and 69 inches.
 - b. Would you expect to meet many Asian adult males over 72 inches? Justify your answer.
 - c. 20\% of Asian adult males are shorter than what value?
 - d. 40\% of Asian adult males are taller than what value?

Click to see Answer

- a. 0.5404
- b. No, because the probability an Asian adult male is taller than 72 inches is only 0.0083
- c. 63.9 inches
- d. 66.63 inches

8. IQ is normally distributed with a mean of 100 and a standard deviation of 15. Suppose one individual is randomly chosen.

- a. Find the probability that a randomly chosen person has an IQ greater than 120.
- b. Find the probability that a randomly chosen person has an IQ less than 85.
- c. MENSA is an organization whose members have the top 2\% of all IQs. Find the minimum IQ needed to qualify for the MENSA organization.

Click to see Answer

- a. 0.0918
- b. 0.1587
- c. 130.81

9. The percent of fat calories that a person in America consumes each day is normally distributed with a mean of 36 and a standard deviation of 10.
 - a. Find the probability that the percent of fat calories a person consumes is more than 40.
 - b. Find the probability that the percent of fat calories a person consumes is between 50 and 60.
 - c. 25\% of people consume less than what percent of fat calories?
 - d. 5\% of people consume more than what percent of fat calories?

Click to see Answer

- a. 0.3446
- b. 0.0726
- c. 29.26\%
- d. 52.45\%

10. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet.
 - a. What is the probability that a random fly ball travels fewer than 220 feet?
 - b. Would you expect many fly balls to travel more than 400 feet? Justify your answer.
 - c. What percentage of fly balls travel between 300 and 350 feet?
 - d. 80\% of fly balls travel for less than what distance?
 - e. 30\% of fly balls travel for more than what distance?

Click to see Answer

- a. 0.2743
- b. No, because the probability that a fly ball travels more than 400 feet is only 0.0013
- c. 0.1359
- d. 292.08 feet
- e. 276.22 feet

11. Suppose that the duration of a particular type of criminal trial is known to be normally distributed with a mean of 21 days and a standard deviation of 7 days.
- Find the probability that a randomly selected trial lasted at least 24 days.
 - Find the probability that a randomly selected trial lasted between 10 and 30 days.
 - Find the probability that a randomly selected trial lasted at most 5 days.
 - 60\% of all trials of this type are completed within how many days?
 - 15\% of all trials of this type take longer than how many days?

Click to see Answer

- 0.3341
 - 0.8427
 - 0.0111
 - 22.77 days
 - 28.26 days
12. Terri Vogel, an amateur motorcycle racer, averages 129.71 seconds per 2.5 mile lap (in a seven-lap race) with a standard deviation of 2.28 seconds. The distribution of her race times is normally distributed.
- Find the percent of her laps that are completed in less than 130 seconds.
 - Find the percent of her laps that are completed in more than 137 seconds.
 - Find the percent of her laps that are completed between 125 and 135 seconds.
 - The fastest 3\% of her laps are under what value?
 - The slowest 5\% of her laps are more than what value?

Click to see Answer

- 0.5506
- 0.0007
- 0.9704
- 125.42 seconds
- 133.46 seconds

to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

PART VI

THE CENTRAL LIMIT THEOREM AND SAMPLING DISTRIBUTIONS

In the previous chapters, we looked at calculating probabilities for individual members from a population. For example, finding the probability that a randomly selected person is taller than 180 cm. But, we are typically less concerned about studying individual elements of a population. Instead, we are more interested in a sample taken from a population under study. So, we are probably more interested in finding the probability that the mean height of a sample of individuals is more than 180 cm, rather than the probability that an individual's height is more than 180 cm.

Why are we concerned about studying a sample? Why are probabilities associated with sample statistics so important? In statistics, we want to study and analyze data about a population so that we can draw conclusions about that population. But, populations are generally very large, and it costs time and money to gather data about populations. Instead, we take a random sample from the population, study and analyze the sample, and use the results from the sample to draw conclusions about the wider population. This process is called **statistical inference**. In order for this process to work correctly and give us reliable conclusions about the population, we have to calculate probabilities associated with sample statistics, such as sample means or sample proportions.

In this chapter, we will study sample means, sample proportions, and their relationship to the **central limit theorem**. The **central limit theorem** is one of the most powerful and useful ideas in all of statistics. The central limit theorem basically says that if we collect samples of size n from a population with mean μ and standard deviation σ , calculate each sample's mean, and create a histogram of those means, then, under the right conditions, the resulting histogram will tend to have an approximate normal distribution. Because the distribution of the sample means follows a normal distribution, under the right conditions, we can use the normal distribution to calculate probabilities about sample means.

CHAPTER OUTLINE

6.1 Sampling Distribution of the Sample Mean

6.2 Sampling Distribution of the Sample Proportion

“6.1 Introduction to Sampling Distributions and the Central Limit Theorem” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

6.1 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

LEARNING OBJECTIVES

- Describe the distribution of the sample mean.
- Solve probability problems involving the distribution of the sample mean.

Suppose all samples of size n are selected from a population with mean μ and standard deviation σ . For each sample, the sample mean \bar{x} is recorded. The probability distribution of these sample means is called **the sampling distribution of the sample means**. The **central limit theorem** describes the properties of the sampling distribution of the sample means.

THE CENTRAL LIMIT THEOREM

Suppose all samples of size n are taken from a population with mean μ and standard deviation σ . The collection of sample means forms a probability distribution called the **sampling distribution of the sample mean**.

1. The mean of the distribution of the sample means, denoted $\mu_{\bar{x}}$, equals the mean of the population.

$$\mu_{\bar{x}} = \mu$$

2. The standard deviation of the sample means (called the standard error of the mean), denoted $\sigma_{\bar{x}}$, equals the standard deviation of the population divided by the square root of the sample size.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

3. The distribution of the sample means follows a normal distribution if **one** of the following conditions is met:
- The population the samples are drawn from is normal, regardless of the sample size n .
 - The sample size $n \geq 30$.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=152#oembed-2>

Video: “Central limit theorem | Inferential statistics | Probability and Statistics | Khan Academy” by Khan Academy [9:49] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=152#oembed-3>

Video: “Sampling distribution of the sample mean | Probability and Statistics | Khan Academy” by Khan Academy [10:52] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=152#oembed-4>

Video: “Standard error of the mean | Inferential statistics | Probability and Statistics | Khan Academy” by Khan Academy [15:15] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Calculating Probabilities for Sample Means

Because the central limit theorem states that the sampling distribution of the sample means follows a normal distribution (under the right conditions), the normal distribution can be used to answer probability questions about sample means. The z -score for the sampling distribution of the sample means is

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where μ is the mean of the population the sample is taken from, σ is the standard deviation of the population the sample is taken from, and n is the sample size.

CALCULATING PROBABILITIES ABOUT SAMPLE MEANS IN EXCEL

Because the sample means follow a normal distribution (under the right conditions), the **norm.dist(x,μ,σ,logic operator)** function can be used to calculate probabilities associated with a sample mean.

- For **x**, enter the value for \bar{x} .
- For **μ**, enter the mean of the sample means μ . Because the mean of the sample means equals the mean of the population the sample is taken from, we enter μ , the mean of the population.
- For **σ**, enter the standard error of the sample means $\frac{\sigma}{\sqrt{n}}$.
- For the logic operator, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

NOTES

1. In this case, we want to calculate probabilities associated with a sample mean. The sample means follow a normal distribution (under the right conditions), which allows us to use the **norm.dist** function to calculate probabilities. Because we are working with sample means, we must enter the **mean** and the **standard distribution** of the **distribution of the sample means** into the **norm.dist** function, and not the mean and standard distribution of the population the samples are taken from. The mean of the sample means equals the mean of the population, so we enter the value of μ into the second field of the **norm.dist** function.
But the standard distribution of the sample means equals $\frac{\sigma}{\sqrt{n}}$, so we must enter this value into the third field of the **norm.dist** function.
2. We use the **norm.dist** function in the same way as we learned previously to calculate the probability a sample mean is less than a given value, a sample mean is greater than a given value, or a sample mean is in between two given values.
3. An alternative approach in Excel is to use the **norm.s.dist(z,true)** function. In the **norm.s.dist** function, we enter the **z**-score for the corresponding value of \bar{x} (using the **z**-score for sample means given above).

EXAMPLE

The length of time, in hours, it takes an “over 40” group of people to play one soccer match is normally distributed with a mean of 2 hours and a standard deviation of 0.5 hours. Suppose a sample of size 25 is drawn randomly from the population.

1. Is the distribution of the sample means normal? Explain.
2. What is the mean and the standard distribution of the distribution of the sample means?
3. What is the probability that the mean of the sample is less than 1.7 hours?
4. What is the probability that the mean of the sample is more than 2.2 hours?
5. What is the probability that the sample mean is between 1.8 hours and 2.3 hours?

Solution

1. Because the population the sample is taken from follows a normal distribution, the distribution of the sample means also follows a normal distribution.
2. The mean of the distribution of the sample means is $\mu_{\bar{x}} = 2$. The standard deviation of the sample means is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{25}} = 0.1$.

3.	Function	norm.dist
	Field 1	1.7
	Field 2	2
	Field 3	0.5/sqrt(25)
	Field 4	true
	Answer	0.0013

The probability the sample mean is less than 1.7 hours is 0.0013 (or 0.13\%).

Note: Because we are calculating a probability for a sample mean, we enter the standard deviation of the sample means 0.5/sqrt(25) into field 3 (and not the standard deviation of the population).

4.

Function	1-norm.dist
Field 1	2.2
Field 2	2
Field 3	0.5/sqrt(25)
Field 4	true
Answer	0.0228

The probability the sample mean is more than **2.2** hours is **0.0228** (or 2.28\%).

5.

Function	norm.dist	-norm.dist
Field 1	2.3	1.8
Field 2	2	2
Field 3	0.5/sqrt(25)	0.5/sqrt(25)
Field 4	true	true
Answer	0.9759	

The probability the sample mean is between **1.8** hours and **2.3** hours is **0.9759** (or 97.59\%).

TRY IT

The length of time taken on the SAT for a group of students has a mean of **2.5** hours and a standard deviation of **0.25** hours. A sample size of **60** is drawn randomly from the population.

1. Is the distribution of the sample means normal? Explain.
2. What is the probability that the sample mean is between **2.4** hours and **2.8** hours?
3. What is the probability that the sample mean is at least **2.6** hours?
4. What is the probability that the sample mean is at most **2.45** hours?

Click to see Solution

1. The distribution of the sample means is normal because the sample size of **60** is greater than **30**.

2.

Function	norm.dist	-norm.dist
Field 1	2.8	2.4
Field 2	2.5	2.5
Field 3	$0.25/\sqrt{60}$	$0.25/\sqrt{60}$
Field 4	true	true
Answer	0.9990	

3.

Function	1-norm.dist
Field 1	2.6
Field 2	2.5
Field 3	$0.25/\sqrt{60}$
Field 4	true
Answer	0.0010

4.

Function	norm.dist
Field 1	2.45
Field 2	2.5
Field 3	$0.25/\sqrt{60}$
Field 4	true
Answer	0.0607

EXAMPLE

In a recent study reported on Oct. 29, 2012, on the Flurry Blog, the mean age of tablet users is **34** years, and the standard deviation is **15** years. Suppose a sample of **100** tablet users is taken.

1. What are the mean and standard deviation for the sample mean ages of tablet users?
2. What is the distribution of the sample means? Explain.
3. Find the probability that the sample mean age is more than **30** years.

Solution

1. The mean of the distribution of the sample means is $\mu_{\bar{x}} = 34$. The standard deviation of the sample means is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = 1.5$.
2. The distribution of the sample means is normal because the sample size of **100** is greater than **30**.

3.

Function	1-norm.dist
Field 1	30
Field 2	34
Field 3	15/sqrt(100)
Field 4	true
Answer	0.9962

The probability the sample mean is more than **30** years of age is **0.9962** (or 99.62\%).

TRY IT

In an article on Flurry Blog, a gaming marketing gap for men between the ages of 30 and 40 is identified. You are researching a start-up game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

Click to see Solution

You need to determine the probability for men whose mean age is between 29 and 35 years of age wanting to play a strategy game.

Function	norm.dist	-norm.dist
Field 1	35	29
Field 2	28	28
Field 3	$4.8/\sqrt{100}$	$4.8/\sqrt{100}$
Field 4	true	true
Answer	0.0186	

There is 1.86% chance that the mean age of men who will play your game is between 29 years and 35 years. Because this is a very low probability, you should not continue your development strategy.

EXAMPLE

The mean number of minutes for app engagement by a tablet user is 8.2 minutes with a standard deviation of 1 minute. Suppose a sample of 60 tablet users is taken.

1. Is the distribution of the sample mean normal? Explain.
2. What are the mean and standard deviation for the sample mean number of minutes for app engagement?
3. Find the probability that the sample mean is between 8 minutes and 8.5 minutes.
4. Find the probability that the sample mean is less than 8.3 minutes.

Solution

1. Because the sample size of 60 is greater than 30, the distribution of the sample means also follows a normal distribution.
2. The mean of the distribution of the sample means is $\mu_{\bar{x}} = 8.2$. The standard deviation of the sample means is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{60}} = 0.13$.

3.

Function	norm.dist	-norm.dist
Field 1	8.5	8
Field 2	8.2	8.2
Field 3	1/sqrt(60)	1/sqrt(60)
Field 4	true	true
Answer	0.9293	

The probability that the sample mean is between 8 and 8.5 minutes is 0.9293 (or 92.93%).

4.

Function	norm.dist
Field 1	8.3
Field 2	8.2
Field 3	1/sqrt(60)
Field 4	true
Answer	0.7807

The probability that the sample mean is less than 8.3 minutes is **0.7807** (or 78.07\%).

TRY IT

Cans of a cola beverage claim to contain **16** ounces with a standard deviation of **0.143** ounces. The amounts in a sample of **34** cans are measured, and the mean is **16.01** ounces. Find the probability that a sample of **34** cans will have an average amount greater than **16.01** ounces. Do the results suggest that cans are filled with an amount greater than **16** ounces?

Click to see Solution

Function	1-norm.dist
Field 1	16.01
Field 2	16
Field 3	0.143/sqrt(34)
Field 4	true
Answer	0.3417

Because there is a 34.17\% probability that the average sample volume is greater than **16.01** ounces,

we should be skeptical of the company's claimed volume. That is, based on this sample, it is likely that the average volume of the cans is higher than the claimed 16 ounces.

As consumers, we would be glad if the average was higher than 16 ounces because we are likely receiving more cola in the can than what we paid for. As the manufacturer, we would need to inspect our bottling process to determine if the process is working within acceptable limits.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=152#oembed-1>

Video: “Excel Statistics 76: Sampling Distribution Of Sample Mean & Central Limit Theorem” by excelisfun [24:06] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately 4 hours each, with a standard deviation of 1.2 hours. Assume the time it takes her to complete one review is normally distributed. Suppose 16 reviews are selected at random.
 - a. What is the mean and standard deviation of the population?
 - b. What is the distribution of the sample means? Explain.
 - c. What is the mean and standard deviation of the sample means?
 - d. Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours.
 - e. Find the probability that the **mean** of a month's reviews will take Yoonie from 3.5 to 4.25 hours.
 - f. Why are the probabilities in (d) and (e) different?
 - g. Find the probability that the mean of a month's reviews will take Yoonie more than 5 hours.

Click to see Answer

- a. ~~4~~ 1.2
 - b. Normal because the population the sample is taken from is normal.
 - c. ~~4~~ 0.3
 - d. 0.2441
 - e. 0.7499
 - f. Part (d) is the probability of a single element from the population, and part (e) is the probability of the sample mean.
 - g. 0.0004
2. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.
- a. What is the probability that the 49 balls travelled an average of less than 230 feet?
 - b. What is the probability that the 49 balls travelled an average of 245 feet to 255 feet?
 - c. What is the probability that the 49 balls travelled an average of more than 260 feet?

Click to see Answer

- a. 0.0026
 - b. 0.5161
 - c. 0.0808
3. According to the CRA, the average length of time for an individual to complete (keep records for, learn, prepare, copy, assemble, and send) their tax return is 10.53 hours with a standard deviation of 2 hours. Suppose we randomly sample 36 taxpayers.
- a. What is the distribution of the sample means? Explain.
 - b. Find the probability that the 36 taxpayers in the sample finished their tax returns in an average of less than 10 hours.
 - c. Would you be surprised if the 36 taxpayers finished their tax returns in an average of more than 11.5 hours? Explain.
 - d. Would you be surprised if one taxpayer finished their tax return in more than 11.5 hours? Explain.

Click to see Answer

- a. Normal because the sample size (36) is greater than 30.
- b. 0.0559
- c. Yes, because the probability the average time is greater than 11.5 hours is only 0.0018.
- d. No, because the probability that an individual taxpayer took more than 11.5 hours is 0.3138.

4. Suppose that a category of world-class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races.
- Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
 - Find the probability that the runner will average less than 140 minutes in these 49 marathons.
 - Find the probability that the runner will average more than 148 minutes in these 49 marathons.

Click to see Answer

- 0.6247
 - 0.0062
 - 0.0668
5. In 1940, the average size of a U.S. farm was 174 acres, and the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940.
- What is the distribution of the sample means? Explain.
 - What is the mean and standard deviation of the sample means?
 - What is the probability that the sample mean is less than 170 acres?
 - What is the probability that the sample mean is more than 180 acres?
 - What is the probability that the sample mean is between 165 and 175 acres?

Click to see Answer

- Normal because the sample size (38) is greater than 30.
 - 174, 8.92
 - 0.3270
 - 0.2506
 - 0.3881
6. The percent of fat calories that a person in America consumes each day is normally distributed with a mean of 36 and a standard deviation of 10. Suppose that 16 individuals are randomly chosen.
- What is the distribution of the sample means?
 - What is the mean and standard deviation of the sample means?
 - Find the probability that the average percent of fat calories consumed in the group of 16 is more than 35.
 - Find the probability that the average percent of fat calories consumed in the group of 16

is less than 30.

- e. Find the probability that the average percent of fat calories consumed in the group of 16 is between 40 and 45.

Click to see Answer

- a. Normal because the population the sample is taken from is normal.
 - b. 36, 2.5
 - c. 0.6554
 - d. 0.0082
 - e. 0.0546
7. The distribution of income in some Third World countries is considered wedge-shaped (many very poor people, very few middle-income people, and even fewer wealthy people). Suppose we pick a country with a wedge-shaped distribution. Suppose the average salary is \$2,000 per year with a standard deviation of \$8,000. We randomly survey 1,000 residents of that country.
- a. How is it possible for the standard deviation to be greater than the average?
 - b. What is the distribution of the sample means? Explain
 - c. Is it likely that the average salary of the 1,000 residents is more than \$2,800? Explain.
 - d. Is it likely that the average salary of the 1,000 residents is less than \$1,800? Explain.
 - e. Why is it more likely that the average salary of the 1,000 residents will be from \$2,000 to \$2,100 than from \$2,100 to \$2,200?

Click to see Answer

- a. Because there are many more poor people compared to middle-income or wealthy people, the mean will be closer to the income of the poorer people. The middle income and wealthy people will have a very large dispersion away from this mean, which causes a large standard deviation.
 - b. Normal because the population the sample size (1,000) is greater than 30.
 - c. No, because the probability that the average salary is more than \$2,800 is only 0.0008.
 - d. Yes, because the probability that the average salary is less than \$1,800 is 0.0008.
 - e. The probability that the average is between \$2,000 and \$2,100 is higher at 0.2146, compared to the probability that the average is between \$2,100 and \$2,200 at 0.1317.
8. NeverReady Batteries has engineered a newer, longer-lasting AAA battery. The company claims this battery has an average life span of 17 hours with a standard deviation of 0.8 hours. Your statistics class questions this claim. As a class, you randomly select 30 batteries and find that the sample mean life span is 16.7 hours. If the process is working properly,

what is the probability of getting a random sample of 30 batteries in which the sample mean lifetime is 16.7 hours or less? Is the company's claim reasonable?

Click to see Answer

0.02. The claim is not reasonable because the probability of getting a sample mean of 16.7 or less is only 0.02. If the population mean was 17, we would expect the probability of a sample mean of 16.7 or less to be higher than 0.02.

9. Your company has a contract to perform preventive maintenance on thousands of air-conditioners in a large city. Based on service records from previous years, the time that a technician spends servicing a unit averages one hour with a standard deviation of one hour. In the coming week, your company will service a simple random sample of 70 units in the city. You plan to budget an average of 1.1 hours per technician to complete the work. Will this be enough time?

Click to see Answer

Yes, because the probability that the average time is less than 1.1 hours is 0.7986.

“6.2 Sampling Distribution of the Sample Mean” and “6.4 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

6.2 SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION

LEARNING OBJECTIVES

- Describe the distribution of the sample proportion.
- Solve probability problems involving the distribution of the sample proportion.

The Central Limit Theorem tells us that the distribution of the sample means follow a normal distribution under the right conditions, which allows us to answer probability questions about the sample mean \bar{x} . Now, we want to investigate the sampling distribution for another important parameter—the sampling distribution of the sample proportion. Once we know what distribution the sample proportions follow, we can answer probability questions about sample proportions.

A **proportion** is the percent, fraction, or ratio of a sample or population that have a characteristic of interest. The **population proportion** is denoted by p , and the **sample proportion** is denoted by \hat{p} .

$$\begin{aligned}\text{Proportion} &= \frac{\text{Number of Items with Characteristic of Interest}}{\text{Total Number of Items}} \\ &= \frac{x}{n}\end{aligned}$$

If the random variable is discrete, such as for categorical data, then the parameter we wish to estimate is the population proportion. This is, of course, the probability of drawing a success in any one random draw. Because we are interested in the number of successes, we are dealing with the binomial distribution. The random variable X is the number of successes, and the parameter we

wish to know is p , the probability of drawing a success, which is the proportion of successes in the population. What is the distribution of the sample proportion \hat{p} ?

THE CENTRAL LIMIT THEOREM FOR SAMPLE PROPORTIONS

Suppose all samples of size n are taken from a population with proportion p . The collection of sample proportions forms a probability distribution called the **sampling distribution of the sample proportion**.

1. The mean of the distribution of the sample proportions, denoted $\mu_{\hat{p}}$, equals the population proportion.

$$\mu_{\hat{p}} = p$$

2. The standard deviation of the sample proportions (called the standard error of the proportion), denoted $\sigma_{\hat{p}}$, is

$$\sigma_{\hat{p}} = \sqrt{\frac{p \times (1 - p)}{n}}$$

3. The distribution of the sample proportions is:
 - Normal if $n \times p \geq 5$ and $n \times (1 - p) \geq 5$.
 - Binomial if one of $n \times p < 5$ or $n \times (1 - p) < 5$.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=154#oembed-1>

Video: “Sampling distribution of sample proportion part 1 | AP Statistics | Khan Academy” by Khan

Academy [9:57] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=154#oembed-2>

Video: “Sampling distribution of sample proportion part 2 | AP Statistics | Khan Academy” by Khan Academy [4:34] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Calculating Probabilities for Sample Proportions using the Normal Distribution

When $n \times p \geq 5$ **and** $n \times (1 - p) \geq 5$, the central limit theorem states that the sampling distribution of the sample proportions follows a normal distribution. In this case, the normal distribution can be used to answer probability questions about sample proportions, and the z -score for the sampling distribution of the sample proportions is

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \times (1-p)}{n}}}$$

where p is the population proportion and n is the sample size.

CALCULATING PROBABILITIES ABOUT SAMPLE PROPORTIONS IN EXCEL USING THE NORMAL DISTRIBUTION

When the distribution of the sample proportions follows a normal distribution (when $n \times p \geq 5$ and $n \times (1 - p) \geq 5$), the **norm.dist(x,μ,σ,logic operator)** function can be used to calculate probabilities associated with a sample proportion.

- For **x**, enter the value for \hat{p} .
- For **μ**, enter the mean of the sample proportions p . Because the mean of the sample proportions equals the proportion of the population the sample is taken from, we enter p , the population proportion.
- For **σ**, enter the standard error of the sample proportions $\sqrt{\frac{p \times (1 - p)}{n}}$.
- For the logic operator, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

NOTES

- In this case, we want to calculate probabilities associated with a sample proportion. The sample proportions follow a normal distribution (under the right conditions), which allows us to use the **norm.dist** function to calculate probabilities. Because we are working with sample proportions, we must enter the **mean** and the **standard distribution** of the **distribution of the sample proportions** into the **norm.dist** function. The mean of the sample proportions equals the population proportion, so we enter the value of p into the second field of the **norm.dist** function. But the standard distribution of the sample proportion is $\sqrt{\frac{p \times (1 - p)}{n}}$, so we must enter this value into the third field of the **norm.dist** function.
- We use the **norm.dist** function in the same way as we learned previously to calculate the probability a sample proportion is less than a given value, a sample proportion is greater than

a given value, or a sample proportion is in between two given values.

- An alternative approach in Excel is to use the **norm.s.dist(z,true)** function. In the **norm.s.dist** function, we enter the z -score for the corresponding value of \hat{p} (using the z -score for sample proportions given above).

EXAMPLE

A recent study asked working adults if they worked most of their time remotely. The study found that 30% of employees spend the majority of their time working remotely. Suppose a sample of 150 working adults is taken.

1. What is the distribution of the sample proportion? Explain.
2. What is the mean and standard deviation of the sample proportion?
3. What is the probability that at most 27% of the workers in the sample work remotely most of the time?
4. What is the probability that at least 51 of the workers in the sample work remotely most of the time?
5. What is the probability that between 32% and 35% of the workers in the sample work remotely most of the time?

Solution

1. $n = 150$ and $p = 0.3$. Checking $n \times p$ and $n \times (1 - p)$:

$$n \times p = 150 \times 0.3 = 45 \geq 5$$

$$n \times (1 - p) = 150 \times (1 - 0.3) = 105 \geq 5$$

Because both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$, the distribution of the sample proportion is normal.

2. The mean of the distribution of the sample proportions is $\mu_{\hat{p}} = 0.3$. The standard deviation of the sample proportions is $\sigma_{\hat{p}} = \sqrt{\frac{p \times (1 - p)}{n}} = \sqrt{\frac{0.3 \times (1 - 0.3)}{150}} = 0.0374$.

3.

Function	norm.dist
Field 1	0.27
Field 2	0.3
Field 3	sqrt(0.3*(1-0.3)/150)
Field 4	true
Answer	0.2113

The probability the sample proportion is at most 27% is **0.2113** (or 21.13%).

Note: Because we are calculating a probability for a sample proportion, we enter the mean of the sample proportions 0.3 (which is the population proportion) into field 2 and the standard deviation of the sample proportions sqrt(0.3*(1-0.3)/150) into field 3.

4. In this case, 51 is not a proportion. It is the number of items in the sample that have the characteristic of interest. We need to convert this 51 out of 150 into a percent: $\frac{51}{150} = 0.34$. This question is asking us to find the probability that at least 34% of the workers in the sample work remotely most of the time.

Function	1-norm.dist
Field 1	0.34
Field 2	0.3
Field 3	$\text{sqrt}(0.3*(1-0.3)/150)$
Field 4	true
Answer	0.1425

The probability the sample proportion is at least 34\% is **0.1425** (or 14.25\%).

5.

Function	norm.dist	-norm.dist
Field 1	0.35	0.32
Field 2	0.3	0.3
Field 3	$\text{sqrt}(0.3*(1-0.3)/150)$	$\text{sqrt}(0.3*(1-0.3)/150)$
Field 4	true	true
Answer	0.2058	

The probability the sample proportion is between 32\% and 35\% is **0.2058** (or 20.58\%).

TRY IT

According to a recent study, 17.5\% of the adult population of Canada are smokers. Suppose a random sample of **200** adult Canadians is taken.

1. What is the distribution of the sample proportion? Explain.
2. What is the mean and standard deviation of the sample proportion?
3. What is the probability that less than **32** of the adults in the sample are smokers?
4. What is the probability that more than 20\% of the adults in the sample are smokers?

5. What is the probability that between 34 and 44 of the adults in the sample are smokers?

Click to see Solution

1. Because $n \times p = 200 \times 0.175 = 35 \geq 5$ and $n \times (1 - p) = 200 \times (1 - 0.175) = 165 \geq 5$ the distribution of the sample proportions is normal.
2. The mean of the distribution of the sample proportions is $\mu_{\hat{p}} = 0.175$. The standard deviation of the sample proportions is

$$\sigma_{\hat{p}} = \sqrt{\frac{p \times (1 - p)}{n}} = \sqrt{\frac{0.175 \times (1 - 0.175)}{200}} = 0.02687.$$

3.

Function	norm.dist
Field 1	0.16
Field 2	0.175
Field 3	sqrt(0.175*(1-0.175)/200)
Field 4	true
Answer	0.2883

4.

Function	1-norm.dist
Field 1	0.2
Field 2	0.175
Field 3	sqrt(0.175*(1-0.175)/200)
Field 4	true
Answer	0.1761

5.

Function	norm.dist	-norm.dist
Field 1	0.22	0.17
Field 2	0.175	0.175
Field 3	sqrt(0.175*(1-0.175)/200)	sqrt(0.175*(1-0.175)/200)
Field 4	true	true
Answer	0.9530	

Calculating Probabilities for Sample Proportions using the Binomial Distribution

When one of $n \times p < 5$ or $n \times (1 - p) < 5$, the sampling distribution of the sample proportions follows a binomial distribution, and so we must use the binomial distribution to answer probability questions about sample proportions. In these cases, we are actually answering probability questions about the number of items with the characteristic of interest, x , not about the sample proportion \hat{p} . In other words, we are answering questions about the number of successes x we get in n trials (the sample size) where the probability of success is the population proportion p . These are exactly the same type of questions we answered previously with the binomial distribution.

CALCULATING PROBABILITIES ABOUT SAMPLE PROPORTIONS IN EXCEL USING THE BINOMIAL DISTRIBUTION

When the distribution of the sample proportions follows a binomial distribution (when one of $n \times p < 5$ or $n \times (1 - p) < 5$), the **binom.dist(x,n,p,logic operator)** function can be used to calculate probabilities associated with a sample proportion.

- For **x**, enter the number of items with the characteristic of interest x .
- For **n**, enter the sample size n . The sample size is the number of trials in the binomial experiment.
- For **p**, enter the population proportion p . The population proportion is the probability of success.
- For the logic operator, enter **true**. **Note:** Because probabilities for sample proportions are generally inequalities ($<$, \leq , $>$, \geq), we enter true for the logic operator. We would only enter false in the case that the probability of the sample proportion exactly equals a given value.

NOTE

We use the **binom.dist** function in the same way as we learned previously to calculate the probability a sample proportion is less than a given value, a sample proportion is at most a given value, a sample proportion is greater than a given value, or a sample proportion is at least a given value.

EXAMPLE

At the local humane society, 3% of the dogs have heartworm disease. Suppose a sample of 60 dogs at the humane society is taken.

1. What is the distribution of the sample proportion? Explain.
2. What is the probability that at most 5% of the dogs in the sample have heartworm disease?
3. What is the probability that less than 7 of the dogs in the sample have heartworm disease?
4. What is the probability that more than 8% of the dogs in the sample have heartworm disease?
5. What is the probability that at least 6 of the dogs in the sample have heartworm disease?

Solution

1. Because $n \times p = 60 \times 0.03 = 1.8 < 5$, the distribution of the sample proportions is binomial.
2. We want to find $P(\hat{p} \leq 0.05)$. Because we are using the binomial distribution, we have to convert 5% into the number of items x in the sample with the required characteristic: $x = 0.05 \times 60 = 3$. In terms of the binomial distribution, we need to find $P(x \leq 3)$.

Function	binom.dist
Field 1	3
Field 2	60
Field 3	0.03
Field 4	true
Answer	0.8943

The probability that at most 5% of the dogs in the sample have heartworm disease is **0.8943** (or 89.43%).

3. We want to find $P(x < 7)$. Because we are using the binomial distribution, this probability is the same as $P(x \leq 6)$.

Function	binom.dist
Field 1	6
Field 2	60
Field 3	0.03
Field 4	true
Answer	0.9979

The probability that less than 7 of the dogs in the sample have heartworm disease is **0.9979** (or 99.79%).

4. We want to find $P(\hat{p} > 0.08)$. Because we are using the binomial distribution, we have to convert 8% into the number of items x in the sample with the required characteristic:
 $x = 0.08 \times 60 = 4.8$. In terms of the binomial distribution, we need to find $P(x > 4.8)$.
 . This is the same as $1 - P(x \leq 4)$.

Function	1-binom.dist
Field 1	4
Field 2	60
Field 3	0.03
Field 4	true
Answer	0.0340

The probability that more than 8\% of the dogs in the sample have heartworm disease is **0.0340** (or 3.4\%).

5. We want to find $P(x \geq 6)$. Because we are using the binomial distribution, this probability is the same as $1 - P(x \leq 5)$.

Function	1-binom.dist
Field 1	5
Field 2	60
Field 3	0.03
Field 4	true
Answer	0.0091

The probability that at least 6 of the dogs in the sample have heartworm disease is **0.0091** (or 0.91\%).

TRY IT

During the past tax season, 92\% of tax returns were filed using an electronic filing system. Suppose a sample of 40 tax returns are selected.

1. What is the distribution of the sample proportions?
2. What is the probability at most 35 of the tax returns in the sample were filed electronically?
3. What is the probability less than 93\% of the tax returns in the sample were filed electronically?
4. What is the probability more than 36 of the tax returns in the sample were filed electronically?
5. What is the probability at least 88\% of the tax returns in the sample were filed electronically?

Click to see Solution

1. Because $n \times (1 - p) = 40 \times (1 - 0.92) = 3.2 < 5$, the distribution of the sample proportions is binomial.

2.

Function	binom.dist
Field 1	35
Field 2	40
Field 3	0.92
Field 4	true
Answer	0.2132

3.

Function	binom.dist
Field 1	37
Field 2	40
Field 3	0.92
Field 4	true
Answer	0.6306

4.

Function	1-binom.dist
Field 1	36
Field 2	40
Field 3	0.92
Field 4	true
Answer	0.6007

5.

Function	1-binom.dist
Field 1	33
Field 2	40
Field 3	0.92
Field 4	true
Answer	0.9624



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=154#oembed-3>

Video: “Excel Statistics 79: Proportions Sampling Distribution” by excelisfun [8:56] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. Suppose in a local school district, 53\% of the population favours a charter school for grades K through five. A simple random sample of 300 people in the district are surveyed.
 - a. What is the distribution of the sample proportion? Explain.
 - b. What is the mean and standard deviation of the sample proportion?
 - c. Find the probability that at least 57\% of people in the sample favour a charter school for grades K through 5.
 - d. Find the probability that no more than 140 of people in the sample favour a charter school for grades K through 5.
 - e. Find the probability that between 50\% and 55\% of people in the sample favour a charter school for grades K through 5.

Click to see Answer

- a. Normal because $n \times p = 159 \geq 5$ and $n \times (1 - p) = 141 \geq 5$.
 - b. 0.53, 0.0288
 - c. 0.0825
 - d. 0.014
 - e. 0.6073
2. Four friends, Janice, Barbara, Kathy, and Roberta, decided to carpool together to get to school. Each day, the driver would be chosen by randomly by selecting one of the four names. They carpool to school for 96 days.
 - a. What is the distribution of the sample proportion?
 - b. Find the probability that Janice is the driver at most 18\% of the time.
 - c. Find the probability that Roberta is the driver more than 16 of those 96 days.

- d. Find the probability that Barbara drives between 24 and 30 of those 96 days.
- e. Find the probability that Kathy is the driver at least 30\% of the time.

Click to see Answer

- a. Normal because $n \times p = 24 \geq 5$ and $n \times (1 - p) = 72 \geq 5$.
- b. 0.0566
- c. 0.9703
- d. 0.4214
- e. 0.1289

3. A question is asked of a class of 200 freshmen, and 23\% of the students know the correct answer. Suppose a sample of 50 students is taken.

- a. What is the mean and standard deviation of the distribution of the sample proportions?
- b. What is the distribution of the sample proportions? Explain.
- c. What is the probability that more than 30\% of the students answered correctly?
- d. What is the probability that less than 20\% of the students answered correctly?
- e. What is the probability that between 21\% and 25\% of the students answered correctly?

Click to see Answer

- a. 0.23, 0.0595
- b. Normal because $n \times p = 11.5 \geq 5$ and $n \times (1 - p) = 38.5 \geq 5$.
- c. 0.1198
- d. 0.3071
- e. 0.6097

4. A virus attacks one in three of the people exposed to it. An entire large city is exposed. Suppose a sample of 70 people in the city is taken.

- a. What is the mean and standard deviation of the distribution of the sample proportions?
- b. What is the distribution of the sample proportions? Explain.
- c. What is the probability that between 21 and 40 of the people in the sample were exposed to the virus?
- d. What is the probability that more than 35\% of the people in the sample were exposed to the virus?
- e. What is the probability that less than 25\% of the people in the same were exposed to the virus?

Click to see Answer

- a. 0.3333, 0.0563

- b. Normal because $n \times p = 23.33 \geq 5$ and $n \times (1 - p) = 46.67 \geq 5$.
- c. 0.7229
- d. 0.3837
- e. 0.0696

5. A local charity is running a lottery to raise funds for its operations. The lottery tickets are “scratch-and-reveal your prize” tickets, where one in every eight tickets is a winner. You purchased 60 tickets.
- a. What is the mean and standard deviation of the distribution of the sample proportions?
 - b. What is the distribution of the sample proportions? Explain.
 - c. What is the probability that less than 10 of your tickets are winners?
 - d. What is the probability that at least 20% of your tickets are winners?
 - e. What is the probability that between 15 and 20 of your tickets are winners?

Click to see Answer

- a. 0.125, 0.0427
- b. Normal because $n \times p = 7.5 \geq 5$ and $n \times (1 - p) = 52.5 \geq 5$.
- c. 0.8354
- d. 0.0395
- e. 0.0017

6. A recent study by the CRA showed that eight out of ten tax returns are filed electronically. A CRA worker takes a random sample of 70 tax returns.
- a. What is the mean and standard deviation of the distribution of the sample proportions?
 - b. What is the distribution of the sample proportions? Explain.
 - c. What is the probability that between 85% and 90% of the returns in the sample were filed electronically?
 - d. What is the probability that at most 50 of the returns in the sample were filed electronically?
 - e. What is the probability that more than 65 of the returns in the sample were filed electronically?

Click to see Answer

- a. 0.8, 0.0478
- b. Normal because $n \times p = 56 \geq 5$ and $n \times (1 - p) = 14 \geq 5$.
- c. 0.1296
- d. 0.0365
- e. 0.0036

7. A game is played repeatedly. A player wins 20\% of the time. Suppose a player plays the game 20 times.
- What is the mean and standard deviation of the distribution of the sample proportions?
 - What is the distribution of the sample proportions? Explain.
 - What is the probability that the player wins at most 7 times?
 - What is the probability that the player wins at least 30\% of the time?
 - What is the probability that the player wins less than 15\% of the time?
 - What is the probability that the player wins more than 10 times?

Click to see Answer

- 0.2, 0.0894
- Binomial because $n \times p = 4 < 5$.
- 0.9679
- 0.1958
- 0.2061
- 0.0006

8. A company inspects products coming through its production process and rejects defective products. 10\% of the items are defective. Suppose a sample of 40 items is taken.
- What is the mean and standard deviation of the distribution of the sample proportions?
 - What is the distribution of the sample proportions? Explain.
 - What is the probability that fewer than 7 of the items in the sample are defective?
 - What is the probability that more than 15\% of the items in the sample are defective?
 - What is the probability that at least 3 of the items in the sample are defective?
 - What is the probability that at most 20\% of the items in the sample are defective?
 - What is the probability that fewer than 80\% of the items in the sample are **not** defective?
 - What is the probability that more than 32 of the items in the sample are **not** defective?
 - What is the probability that at least 95\% of the items in the sample are **not** defective?
 - What is the probability that at most 35 of the items in the sample are **not** defective?

Click to see Answer

- 0.1, 0.0474
- Binomial because $n \times p = 4 < 5$.
- 0.9005
- 0.0995
- 0.7772
- 0.9845

- g. 0.0155
- h. 0.9581
- i. 0.2228
- j. 0.3710

9. A recent market research study reported that six out of seven people prefer coffee to tea.

Suppose a random sample of 28 people is taken.

- a. What is the distribution of the sample proportions? Explain.
- b. What is the probability that fewer than 70\% of the people in the sample prefer coffee to tea?
- c. What is the probability that more than 20 of the people in the sample prefer coffee to tea?
- d. What is the probability that at least 90\% of the people in the sample prefer coffee to tea?
- e. What is the probability that at most 22 of the people in the sample prefer coffee to tea?

Click to see Answer

- a. Binomial because $n \times (1 - p) = 4 < 5$.
- b. 0.0131
- c. 0.9622
- d. 0.2158
- e. 0.2020

10. At a certain university, seven out of ten students are enrolled in full-time programs. Suppose a random sample of 15 students is taken.

- a. What is the distribution of the sample proportions? Explain.
- b. What is the probability that at most 60\% of the students in the sample are full-time students?
- c. What is the probability that more than 82\% of the students in the sample are full-time students?
- d. What is the probability that fewer than 7 of the students in the sample are full-time students?
- e. What is the probability that at least 10 of the students in the sample are full-time students?
- f. What is the probability that at most 4 of the students in the sample are **not** full-time students?
- g. What is the probability that at least 35\% of the students in the sample are **not** full-time students?

students?

Click to see Answer

- a. Binomial because $n \times (1 - p) = 4.5 < 5$.
- b. 0.2784
- c. 0.1268
- d. 0.0152
- e. 0.7216
- f. 0.5155
- g. 0.2784

“6.3 Sampling Distribution of the Sample Proportion” and “6.4 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

PART VII

CONFIDENCE INTERVALS FOR SINGLE POPULATION PARAMETERS

When real estate boards post statements about the average rent in a particular city or area, how do they come up with that number? Are they going around to every single renter in that area, asking them what they pay for rent each month, and then calculating the average? In a very, very small town, this might be possible. But in large cities with lots of renters, such an undertaking would be very time consuming and very expensive. Instead, a sample of rents is taken, and the average of the sample is used to estimate the average rent in that city or town. The average from the sample is called a **point estimate** of the actual average rent for the entire population. We would not expect the point estimate to equal that actual average rent, and there is a possibility that the point estimate is not particularly close to the actual average rent. Is there a way to improve the result in order to get a better estimate of the average rent? The answer is yes, by creating an interval estimate.

In the run-up to any election, polls are taken to **estimate** what percentage of the population will vote for the different people or parties involved. Pollsters call a random sample of the population, ask the people in the sample how they plan to vote, and, from that data, create an estimate of what the voting public will do on election day. The percentages obtained from the sample are called **point estimates** of the actual population percent. Because polls are based on sample data and not the entire population, there are gaps between the sample statistic (the point estimate) and the corresponding population parameter. Sometimes, these gaps are quite large, which means the sample statistic is not a good estimate of the population parameter. Instead of relying on just the point estimate, pollsters use the point estimate to create an interval estimate called a **confidence interval**. We can see this confidence interval at work when polls are posted in the media—in the fine print, and there are usually statements saying something like “plus or minus 2.5%, 19 times out of 20”. In other words, the pollsters are saying that there is a 95% probability that the actual percentage of the population that will vote for a particular party falls inside the interval created by adding and subtracting 2.5% from the point estimate.

This type of statistics, using sample data to make generalizations about an unknown population parameter, is called **inferential statistics**. In this case, the type of inferential statistics we are interested in are called **confidence intervals**. A random sample is taken from the population

under study, and the sample statistic is calculated as a point estimate of a population parameter. Because the point estimate is most likely not the exact value of the population parameter, we construct an interval estimate, a **confidence interval**. With a certain probability, we will be able to say that the population parameter is inside the confidence interval.

CHAPTER OUTLINE

7.1 Introduction to Confidence Intervals

7.2 Confidence Intervals for a Single Population Mean with Known Population Standard Deviation

7.3 Confidence Intervals for a Single Population Mean with Unknown Population Standard Deviation

7.4 Confidence Intervals for a Population Proportion

7.5 Calculating the Sample Size for a Confidence Interval

“7.1 Introduction to Confidence Intervals” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

7.1 INTRODUCTION TO CONFIDENCE INTERVALS

LEARNING OBJECTIVES

- Differentiate between point estimates and interval estimates.
- Explain key terms related to confidence intervals.

The marketing department of an entertainment company is interested in the mean number of songs a consumer downloads a month from an online music service. How could the marketing department find this mean? They could take a sample of consumers and calculate the sample mean \bar{x} and the sample standard deviation s of the sample. The sample mean \bar{x} is a **point estimate** for the population mean μ . The sample standard deviation s is a **point estimate** for the population standard deviation σ .

In general, when a sample is taken from a population, the sample statistic calculated from the sample is a **point estimate** for the corresponding population parameter. The point estimate is a single value used to estimate the population parameter. For example, the sample mean \bar{x} is a point estimate for the population mean μ and the sample proportion \hat{p} is a point estimate for the population proportion p .

There are a few issues with relying on just a point estimate to estimate a population parameter. Different samples from the same population will produce different point estimates, and how do we know which point estimate is the best one? Also, a point estimate is a single value, which may or may not be close to the actual population parameter, and how do we know how far away the population parameter is from the point estimate. The difference between the point estimate and the population parameter is called the **sampling error**. Because the population parameter is

unknown, we cannot answer these questions. We have no way to know if a point estimate over- or under-estimates a population parameter or by how much.

Instead of a point estimate, an interval estimate called a confidence interval, provides us with additional insight into estimating a population parameter. Instead of being just one number, a confidence interval is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is **likely** to include the unknown population parameter.

Suppose, for the above example, we do not know the population mean μ , but we do know that the population standard deviation is $\sigma = 1$ and the sample size is $n = 100$. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1$$

The empirical rule, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean \bar{x} will be within two standard deviations of the population mean μ . For our example, two standard deviations is $2 \times 0.1 = 0.2$. The sample mean \bar{x} is likely to be within 0.2 units of μ .

Because \bar{x} is within 0.2 units of μ , which is unknown, μ is likely to be within 0.2 units of \bar{x} in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations ($2 \times 0.1 = 0.2$) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples. Suppose that a sample produced a sample mean $\bar{x} = 2$. Then the unknown population mean μ is between $\bar{x} - 0.2 = 2 - 0.2 = 1.8$ and $\bar{x} + 0.2 = 2 + 0.2 = 2.2$

We say that we are **95% confident** that the (unknown) population mean number of songs downloaded per month is between 1.8 and 2.2. **The 95% confidence interval is the interval with a lower limit 1.8 and upper limit 2.2.**

The 95% confidence interval implies two possibilities. Either the interval 1.8 to 2.2 contains the true mean μ , or our sample produced an \bar{x} that is not within 0.2 units of the true mean μ . Because we are 95% confident that the true population mean is inside the interval, the second possibility is that the population mean is not inside the interval, which happens for only 5% of all the samples.

Remember that a confidence interval is created for an unknown population parameter like the population mean μ . Confidence intervals for some parameters have the form:

$$\text{Lower Limit} = \text{point estimate} - \text{margin of error}$$
$$\text{Upper Limit} = \text{point estimate} + \text{margin of error}$$

Note that the margin of error depends on the confidence level and the standard error of the mean.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=160#oembed-1>

Video: “Understanding Confidence Intervals: Statistics Help” by Dr Nic’s Maths and Stats [4:02] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

“7.1 Introduction to Confidence Intervals” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

7.2 CONFIDENCE INTERVALS FOR A SINGLE POPULATION MEAN WITH KNOWN POPULATION STANDARD DEVIATION

LEARNING OBJECTIVES

- Calculate and interpret confidence intervals for estimating a population mean where the population standard deviation is known.

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. To construct a confidence interval for a single unknown population mean μ , **where the population standard deviation is known**, we need \bar{x} , which is the **point estimate** of the unknown population mean μ .

The confidence interval estimate will have the form:

$$\text{Lower Limit} = \bar{x} - \text{margin of error}$$

$$\text{Upper Limit} = \bar{x} + \text{margin of error}$$

The margin of error depends on the **confidence level**. The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level. Typically, confidence levels of 90\% or higher are used because we want to have a high level of certainty about the results.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=169#oembed-1>

Video: “Confidence Intervals – Introduction” by Joshua Emmanuel [3:35] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

EXAMPLE

Suppose we have collected data from a sample. The sample mean is 7 and the margin of error is 2.5.

The confidence interval is:

$$\text{Lower Limit} = 7 - 2.5 = 4.5$$

$$\text{Upper Limit} = 7 + 2.5 = 9.5$$

If the confidence level is 95%, then we say that “We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5.”

TRY IT

Suppose we have data from a sample. The sample mean is 15 and the margin of error is 3.2. What is the confidence interval estimate for the population mean?

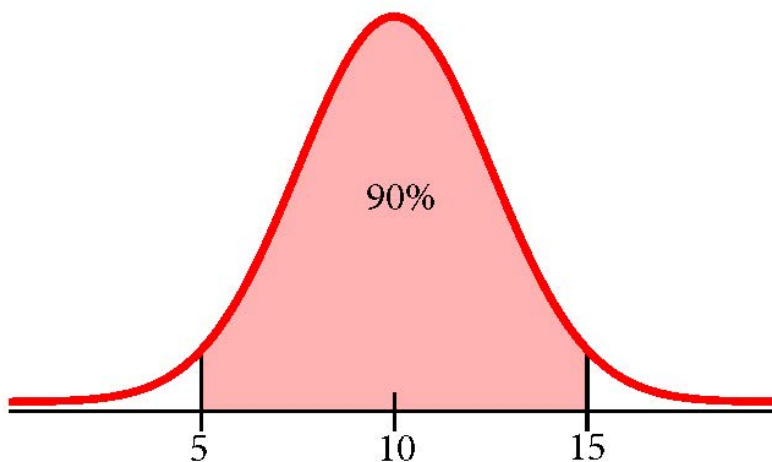
Click to see Solution

$$\text{Lower Limit} = 15 - 3.2 = 11.8$$

$$\text{Upper Limit} = 15 + 3.2 = 18.2$$

A confidence interval for a population mean with a known population standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90% confidence interval with a lower limit of 5 and an upper limit of 15.

To get a 90% confidence interval, we must include the central 90% of the normal distribution. If we include the central 90%, we leave out a total of 10% in both tails or 5% in each tail of the normal distribution.



To capture the central 90%, we must go out 1.645 “standard deviations” on either side of the calculated sample mean. The value 1.645 is the z -score from a standard normal distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the “standard deviation” used is appropriate for the parameter we are estimating. So, in this section, we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$ (the standard deviation of the sample means). The fraction $\frac{\sigma}{\sqrt{n}}$ is commonly called the **standard error of the mean** in order to clearly distinguish the standard deviation for a sample mean from the population standard deviation σ .

Constructing the Confidence Interval

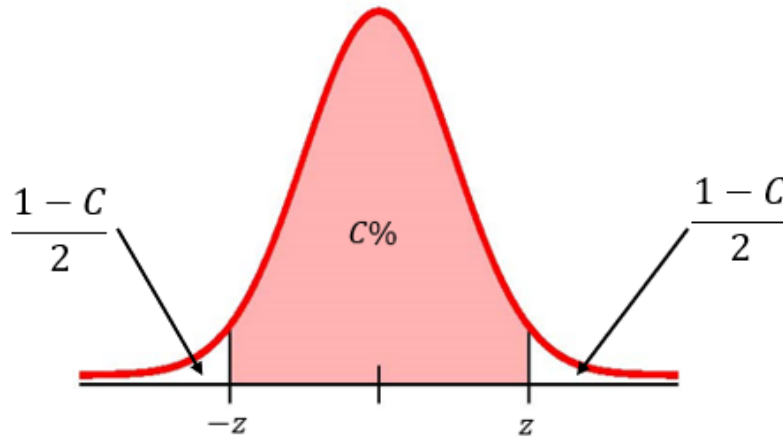
To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

1. Calculate the sample mean \bar{x} from the sample data. Remember, in this section, we already know the population standard deviation σ .
2. Find the z -score that corresponds to the confidence level C .
3. Calculate the limits for the confidence interval.
4. Write a sentence that interprets the estimate in the context of the problem. (Explain what the confidence interval means in the words of the problem.)

We will first examine each step in more detail and then illustrate the entire process with some examples.

Finding the z -score for the Confidence Level

When we know the population standard deviation σ , we use a standard normal distribution to calculate the margin of error and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution. The confidence level C is the area in the middle of the standard normal distribution. The remaining area, $1 - C$, is split equally between the two tails so that each of the tails contains an area equal to $\frac{1 - C}{2}$.



The z -score needed to construct the confidence interval is the z -score so that the **entire** area to the left of z -score equals the area in the middle (the confidence level C) plus the area in the left tail $\left(\frac{1-C}{2}\right)$. That is, the required z -score for the confidence interval is the z -score so that the entire area to the left of the z -score is

$$C + \frac{1-C}{2}$$

For example, if the confidence level is 95%, then the area in the **centre** of the standard normal distribution is 0.95 and the area in the left tail is $\frac{1-0.95}{2} = 0.025$. We would need to find the z -score so that the entire area to the left of the z -score equals $0.95 + 0.025 = 0.975$.

CALCULATING THE z -SCORE FOR A CONFIDENCE INTERVAL IN EXCEL

To find the z -score to construct a confidence interval with confidence level C , use the **norm.s.inv(area to the left of z)** function.

- For **area to the left of z**, enter the **entire** area to the left of the z -score required. For a confidence interval, the area to the left of z is $C + \frac{1-C}{2}$.

The output from the **norm.s.inv** function is the value of z -score needed to construct the confidence interval.

NOTE

The **norm.s.inv** function requires that we enter the **entire** area to the **left** of the unknown z -score. This area includes the confidence level C (the area in the middle of the distribution) plus the remaining area in the left tail $\frac{1 - C}{2}$.

Calculating the Margin of Error

The margin of error for a confidence interval with confidence level C for an unknown population mean μ when the population standard deviation σ is known is

$$\text{Margin of Error} = z \times \frac{\sigma}{\sqrt{n}}$$

where z is the z -score from the standard normal distribution so that the area to the left of z is $C + \frac{1 - C}{2}$.

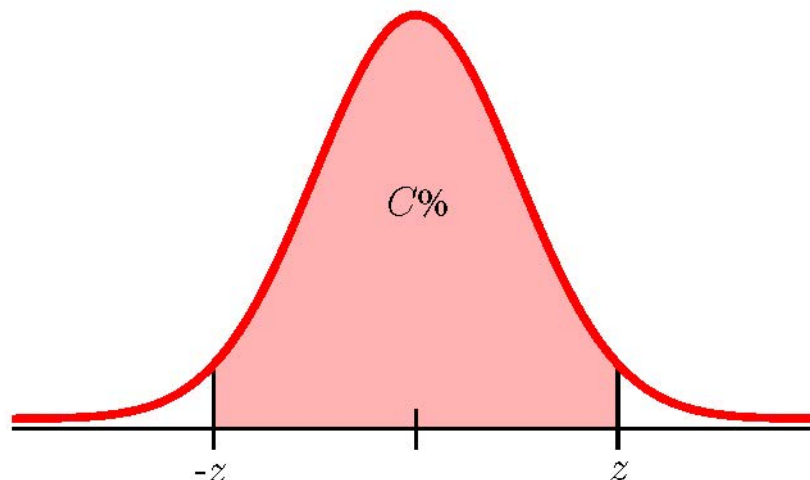
Calculating the Limits of the Confidence Interval

The limits for the confidence interval with confidence level C for an unknown population mean μ when the population standard deviation σ is known are

$$\text{Lower Limit} = \bar{x} - z \times \frac{\sigma}{\sqrt{n}}$$

$$\text{Upper Limit} = \bar{x} + z \times \frac{\sigma}{\sqrt{n}}$$

where z is the z -score from the standard normal distribution so that the area to the left of z is $C + \frac{1 - C}{2}$.



Interpreting a Confidence Interval

The interpretation should clearly state the confidence level C , explain what population parameter is being estimated (in this case, a **population mean**), and state the confidence interval (both endpoints). For example, “We estimate with ____% confidence that the true population mean (include the context of the problem) is between ____ and ____ (include appropriate units).”



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=169#oembed-2>

Video: “Confidence Interval for a population mean – σ known” by Joshua Emmanuel [4:31] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

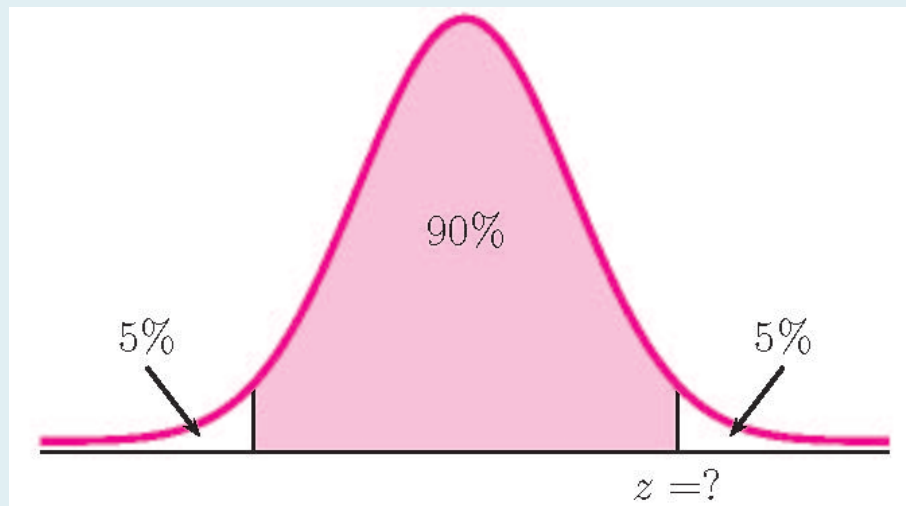
EXAMPLE

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of 3 points. A random sample of 36 scores is taken and has a sample mean of 68 points.

1. Find a 90% confidence interval for the mean exam score.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to conclude that the mean exam score for all the exams is 70? Explain.

Solution

1. To find the confidence interval, we need to find the z -score for the 90% confidence interval. This means that we need to find the z -score from the standard normal distribution so that the entire area to the left of z is $0.9 + \frac{1 - 0.9}{2} = 0.95$.



Function	norm.s.inv
Field 1	0.95
Answer	1.6448...

So $z = 1.6448\dots$. From the question, $\bar{x} = 68$, $\sigma = 3$ and $n = 36$. The 90% confidence interval is

$$\begin{aligned}\text{Lower Limit} &= \bar{x} - z \times \frac{\sigma}{\sqrt{n}} \\ &= 68 - 1.6448\dots \times \frac{3}{\sqrt{36}} \\ &= 67.18\end{aligned}$$

$$\begin{aligned}\text{Upper Limit} &= \bar{x} + z \times \frac{\sigma}{\sqrt{n}} \\ &= 68 + 1.6448\dots \times \frac{3}{\sqrt{36}} \\ &= 68.82\end{aligned}$$

2. We are 90% confident that the mean exam score is between 67.18 points and 68.82 points.
3. It is not reasonable to conclude that the mean exam score is 70 points because 70 points is outside the confidence interval. (In this case, there is a 90% chance that the actual mean exam score is in between 67.18 and 68.82 and only a 10% chance that the mean exam score is outside this interval. So it is unlikely (but not impossible) that the actual mean exam score is a value outside of the confidence interval.)

NOTES

1. When calculating the limits for the confidence interval, keep all of the decimals in the z -score and other values throughout the calculation. This will ensure that there is no round-off error in the answers. Use Excel to do the calculation of the limits, clicking on the cells containing the z -score and any other values, to ensure that all of the decimal places are used in the calculation.
2. When writing down the interpretation of the confidence interval, make sure to include the

confidence level, the actual population mean captured by the confidence interval (i.e. be specific to the context of the question), and appropriate units for the limits.

3. With a confidence level of $C\%$, $C\%$ of all confidence intervals constructed contain the true population parameter. For the above example, this means that 90% of all confidence intervals constructed this way contain the true mean exam score. In other words, if we constructed **100** of these confidence intervals (using **100** different samples of size **36**), we would expect **90** of them to contain the true mean exam score.

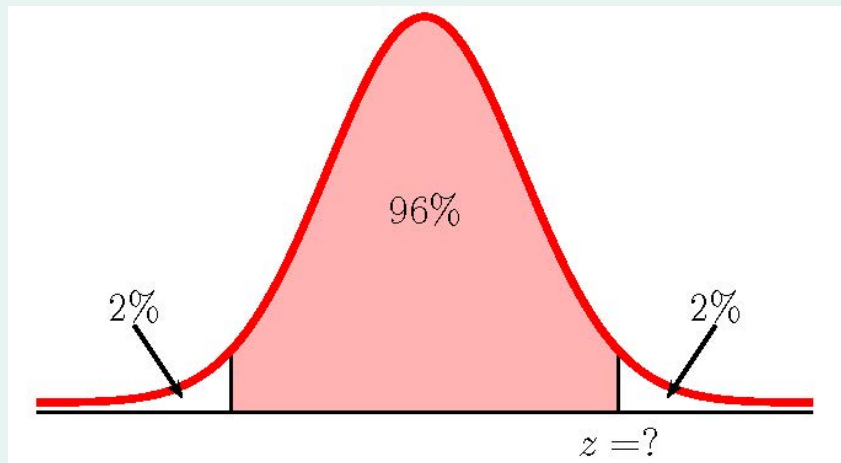
TRY IT

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of **6** minutes. A random sample of **28** pizza delivery restaurants is taken and has a sample mean delivery time of **36** minutes.

1. Find a 96% confidence interval for the mean delivery time.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to claim that the mean delivery time is **35** minutes? Explain.

Click to see Solution

1.	Function	norm.s.inv
	Field 1	0.98
	Answer	2.053...



$$\begin{aligned}\text{Lower Limit} &= \bar{x} - z \times \frac{\sigma}{\sqrt{n}} \\ &= 36 - 2.053 \dots \times \frac{6}{\sqrt{28}} \\ &= 33.67\end{aligned}$$

$$\begin{aligned}\text{Upper Limit} &= \bar{x} + z \times \frac{\sigma}{\sqrt{n}} \\ &= 36 + 2.053 \dots \times \frac{6}{\sqrt{28}} \\ &= 38.05\end{aligned}$$

2. We are 96\% confident that the mean delivery time is between **33.67** minutes and **38.05** minutes.
3. It is reasonable to conclude that the mean delivery time is **35** minutes because **35** minutes is inside the confidence interval.

EXAMPLE

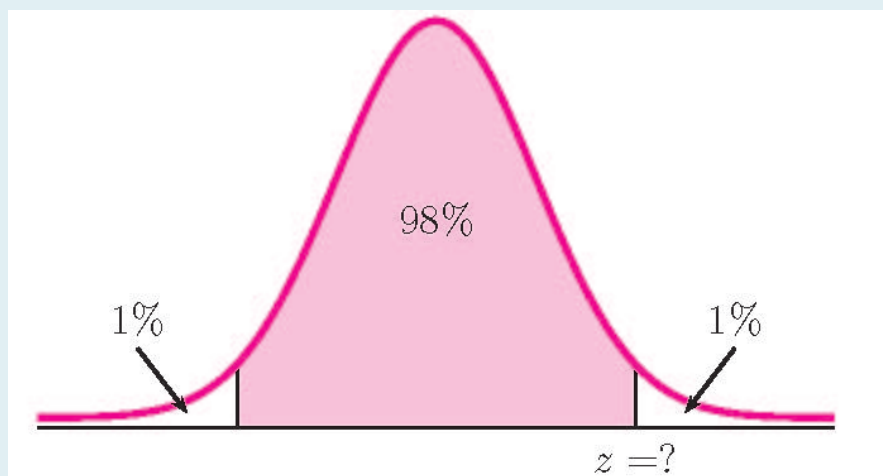
The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. This table shows the highest SAR level for a random selection of cell phone models as measured by the FCC.

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messenger III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

1. Find a 98% confidence interval for the mean of the Specific Absorption Rates (SARs) for cell phones. Assume that the population standard deviation is $\sigma = 0.337$
2. Interpret the confidence interval found in part 1.

Solution

1. To find the confidence interval, we need to find the z -score for the 98% confidence interval. This means that we need to find the z -score from the standard normal distribution so that the entire area to the left of z is $0.98 + \frac{1 - 0.98}{2} = 0.99$.



Function	norm.s.inv
Field 1	0.99
Answer	2.3263...

So $z = 2.3263 \dots$. From the sample data supplied in the question $\bar{x} = 1.0237 \dots$ and $n = 30$. The population standard deviation is $\sigma = 0.337$. The 98% confidence interval is

$$\begin{aligned}
 \text{Lower Limit} &= \bar{x} - z \times \frac{\sigma}{\sqrt{n}} \\
 &= 1.0237 \dots - 2.3263 \dots \times \frac{0.377}{\sqrt{30}} \\
 &= 0.8806
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \bar{x} + z \times \frac{\sigma}{\sqrt{n}} \\
 &= 1.0237 \dots + 2.3262 \dots \times \frac{0.377}{\sqrt{30}} \\
 &= 1.1839
 \end{aligned}$$

- We are 98% confident that the mean of the Specific Absorption Rates is between 0.8806 watts per kilogram and 1.1839 watts per kilogram.

TRY IT

This table shows a different random sampling of 20 cell phone models. Assume that the population standard deviation is $\sigma = 0.337$.

Phone Model	SAR	Phone Model	SAR
Blackberry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

1. Construct a 93\% confidence interval for the mean SAR for cell phones certified for use in the United States.
2. Interpret the confidence interval found in part 1.

Click to see Solution

1.	Function	norm.s.inv
	Field 1	0.965
	Answer	1.8119...

$$\begin{aligned}\text{Lower Limit} &= \bar{x} - z \times \frac{\sigma}{\sqrt{n}} \\ &= 0.94\dots - 1.8119\dots \times \frac{0.337}{\sqrt{20}} \\ &= 0.8035\end{aligned}$$

$$\begin{aligned}\text{Upper Limit} &= \bar{x} + z \times \frac{\sigma}{\sqrt{n}} \\ &= 0.94\dots + 1.8119\dots \times \frac{0.337}{\sqrt{20}} \\ &= 1.0766\end{aligned}$$

2. We are 93\% confident that the mean of the Specific Absorption Rates is between **0.8035** watts per kilogram and **1.0766** watts per kilogram.

NOTE

Notice the difference in the confidence intervals calculated in the Example and Try It just completed. These intervals are different for several reasons: they were calculated from different samples, the samples were different sizes, and the intervals were calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information.

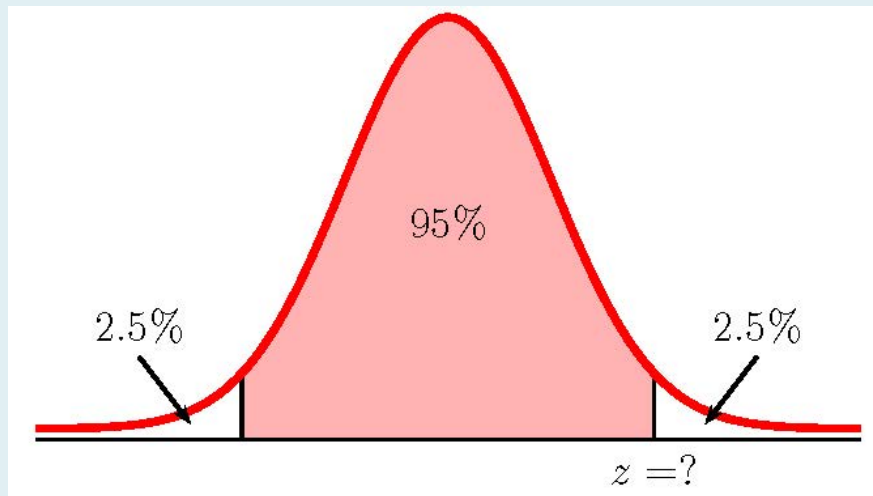
Changing the Confidence Level

EXAMPLE

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of 3 points. A random sample of 36 scores is taken and gives a sample mean of 68 points. Previously, we found a 90% confidence interval for the mean exam score. Now, find a 95% confidence interval for the mean exam score. Interpret the 95% confidence interval.

Solution

To find the confidence interval, we need to find the z -score for the 95% confidence interval. This means that we need to find the z -score from the standard normal distribution so that the entire area to the left of z is $0.95 + \frac{1 - 0.95}{2} = 0.975$.



Function	norm.s.inv
Field 1	0.975
Answer	1.9599...

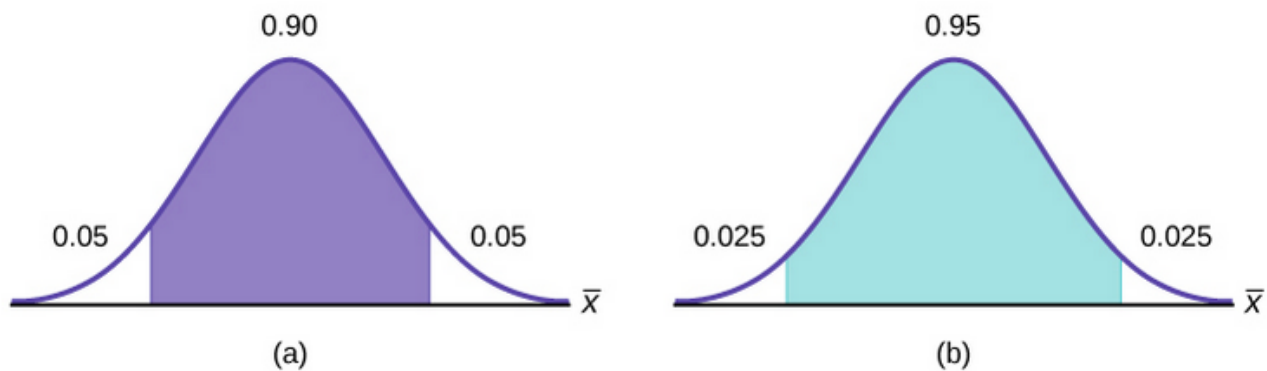
So $z = 1.9599 \dots$. From the question $\bar{x} = 68$, $\sigma = 3$ and $n = 36$. The 95% confidence interval is

$$\begin{aligned}\text{Lower Limit} &= \bar{x} - z \times \frac{\sigma}{\sqrt{n}} \\ &= 68 - 1.9599 \dots \times \frac{3}{\sqrt{36}} \\ &= 67.02\end{aligned}$$

$$\begin{aligned}\text{Upper Limit} &= \bar{x} + z \times \frac{\sigma}{\sqrt{n}} \\ &= 68 + 1.9599 \dots \times \frac{3}{\sqrt{36}} \\ &= 68.98\end{aligned}$$

We are 95% confident that the mean exam score is between 67.02 points and 68.98 points.

For the exam scores examples, the 90% confidence interval has a lower limit of 67.18 and an upper limit of 68.82 and the 95% confidence interval has a lower limit of 67.02 and an upper limit of 68.98. Notice that the 95% confidence interval is wider (the distance between the limits is larger in the 95% confidence interval). If we look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually contains the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.



When the confidence level increases, the margin of error increases, and the effect makes the confidence interval wider. When the confidence level decreases, the margin of error decreases, and the effect makes the confidence interval narrower.

Changing the Sample Size

EXAMPLE

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of 3 points. Previously, we found a 90% confidence interval for the mean exam score using a sample of size 36 with a sample mean of 68.

1. Suppose everything is kept the same, but the sample size is 100 (instead of 36). Find the 90% confidence interval.
2. Suppose everything is kept the same, but the sample size is 25 (instead of 36). Find the 90% confidence interval.

Solution

1.	Function	norm.s.inv
	Field 1	0.95
	Answer	1.6448...

$$\begin{aligned}
 \text{Lower Limit} &= \bar{x} - z \times \frac{\sigma}{\sqrt{n}} \\
 &= 68 - 1.6448 \dots \times \frac{3}{\sqrt{100}} \\
 &= 67.51
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \bar{x} + z \times \frac{\sigma}{\sqrt{n}} \\
 &= 68 + 1.6448 \dots \times \frac{3}{\sqrt{100}} \\
 &= 68.49
 \end{aligned}$$

2.

Function	norm.s.inv
Field 1	0.95
Answer	1.6448...

$$\begin{aligned}
 \text{Lower Limit} &= \bar{x} - z \times \frac{\sigma}{\sqrt{n}} \\
 &= 68 - 1.6448 \dots \times \frac{3}{\sqrt{25}} \\
 &= 67.01
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \bar{x} + z \times \frac{\sigma}{\sqrt{n}} \\
 &= 68 + 1.6448 \dots \times \frac{3}{\sqrt{25}} \\
 &= 69.27
 \end{aligned}$$

For the exam scores examples, the 90\% confidence interval with a sample size of 36 has a lower limit of 67.18 and an upper limit of 68.82, with a sample size of 100 has a lower limit of 67.51 and an upper limit of 68.49, and with a sample size of 25 has a lower limit of 67.01 and an upper limit of 69.27. When the sample size increases, the confidence interval is narrower. When the sample

size decreases, the confidence interval is wider. Generally, the smaller the sample size, the wider the confidence interval needs to be in order to achieve the same level of confidence.

When the sample size increases, the margin of error decreases, and the effect makes the confidence interval narrower. When the sample size decreases, the margin of error increases, and the effect makes the confidence interval wider.

Exercises

1. Statistics Canada conducts a study to determine the time needed to complete the short form of the census. Statistics Canada surveys 200 people and found that the sample mean is 8.2 minutes. The population standard deviation of 2.2 minutes. The population distribution is assumed to be normal.
 - a. Construct a 90\% confidence interval for the mean time to complete the short form of the census.
 - b. Interpret the confidence interval found in part (a).
 - c. Is it reasonable to conclude the mean time to complete the short form is 10 minutes? Explain.
 - d. If Statistics Canada wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?
 - e. If Statistics Canada did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence?

Click to see Answer

- a. Lower Limit = 7.94, Upper Limit = 8.46
 - b. There is a 90\% probability that the mean time to complete the short form of the census is between 7.94 minutes and 8.46 minutes.
 - c. No, because 10 minutes is outside the confidence interval.
 - d. Increase the sample size.
 - e. Confidence level would decrease.
2. A sample of 20 heads of lettuce was selected. The weight of each head of lettuce was then recorded. The mean weight from the sample was 2.2 pounds. The population standard deviation is known to be 0.2 pounds. Assume that the population distribution of head weight is normal.

- a. Construct a 92\% confidence interval for the population mean weight of the heads of lettuce.
- b. Interpret the confidence interval found in part (a).
- c. Construct a 98\% confidence interval for the population mean weight of the heads of lettuce.
- d. In complete sentences, explain why the confidence interval in part (c) is larger than in part (a).
- e. What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?
- f. What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

Click to see Answer

- a. Lower Limit = 2.12, Upper Limit = 2.28
 - b. There is a 92\% probability that the mean weight of heads of lettuce is between 2.12 pounds and 2.28 pounds.
 - c. Lower Limit = 2.10, Upper Limit = 2.30
 - d. For the same sample size and sample data, a higher confidence level requires a wider interval to achieve the required accuracy.
 - e. Confidence level would increase.
 - f. The interval becomes narrower.
3. A college wants to estimate the mean age of students entering the winter semester. Suppose that 75 winter students were randomly selected, and the mean age for the sample was 30.4 years. The population standard deviation of the ages of the students is 15 years.
- a. Construct a 99\% confidence interval for the mean age of winter semester students at the college.
 - b. Interpret the confidence interval found in part (a).
 - c. Is it reasonable for the college to claim that the mean age of its winter semester students is 35? Explain.
 - d. Using the same mean, standard deviation, and level of confidence, suppose that the sample size is 69 instead of 25. Would the error bound become larger or smaller?
 - e. Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90\%?

Click to see Answer

- a. Lower Limit = 25.94, Upper Limit = 34.86
- b. There is a 99\% probability that the mean age of winter semester students is between

25.94 years and 34.86 years.

- c. No, because 35 is outside the confidence interval.
- d. Smaller.
- e. Error bound would get smaller.

4. Suppose that an accounting firm does a study to determine the time needed to complete one person's tax forms. It randomly surveys 100 people and finds that the sample mean is 23.6 hours. The population standard deviation of 7.0 hours. The population distribution is assumed to be normal.

- a. Construct a 97% confidence interval for the mean time to complete the tax forms.
- b. Interpret the confidence interval found in part (a).
- c. Is it reasonable for the firm to claim that the mean time to complete the tax forms is 25 hours? Explain.
- d. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?
- e. If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence?

Click to see Answer

- a. Lower Limit = 22.08, Upper Limit = 25.12
- b. There is a 97% probability that the mean time to complete the tax forms is between 22.08 hours and 25.12 years.
- c. Yes, because 25 is inside the confidence interval.
- d. Increase the sample size.
- e. Decreases.

5. A sample of 16 small bags of the same brand of candies was selected, and the mean weight was 125 grams. Assume that the population distribution of bag weights is normal. The population standard deviation is known to be 6.25 grams.

- a. Construct a 90% confidence interval for the mean weight of the candies.
- b. Construct a 98% confidence interval for the population mean weight of the candies.
- c. In complete sentences, explain why the confidence interval in part (b) is larger than the confidence interval in part (a).
- d. In complete sentences, give an interpretation of what the interval in part (b) means.

Click to see Answer

- a. Lower Limit = 122.43, Upper Limit = 127.57
- b. Lower Limit = 121.37, Upper Limit = 128.63

- c. For the same sample size and sample data, a higher confidence level requires a wider interval to achieve the required accuracy.
 - d. There is a 98\% probability that the mean weight of the candies is between 121.37 grams and 128.63 years.
6. What is meant by the term “90\% confident” when constructing a confidence interval for a mean?
- a. If we took repeated samples, approximately 90\% of the samples would produce the same confidence interval.
 - b. If we took repeated samples, approximately 90\% of the confidence intervals calculated from those samples would contain the sample mean.
 - c. If we took repeated samples, approximately 90\% of the confidence intervals calculated from those samples would contain the true value of the population mean.
 - d. If we took repeated samples, the sample mean would equal the population mean in approximately 90\% of the samples.

Click to see Answer

c

“7.2 Confidence Intervals for a Single Population Mean with Known Population Standard Deviation” and “7.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

7.3 CONFIDENCE INTERVALS FOR A SINGLE POPULATION MEAN WITH UNKNOWN POPULATION STANDARD DEVIATION

LEARNING OBJECTIVES

- Calculate and interpret confidence intervals for estimating a population mean where the population standard deviation is unknown.

In practice, we rarely know the population standard deviation. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ , and proceeded as before to calculate a confidence interval with close enough results. However, statisticians ran into problems when the sample size was small because a small sample size created inaccuracies in the confidence interval.

William S. Goset (1876–1937) of the Guinness Brewery in Dublin, Ireland, ran into this problem. His experiments with hops and barley produced very few samples. Just replacing the population standard deviation σ with the sample standard deviation s did not produce accurate results when he tried to calculate a confidence interval. Goset realized that he could not use a normal distribution for the calculation. He found that the actual distribution depends on the sample size. This problem led him to “discover” what is called the **Student’s t -distribution**. The name comes from the fact that Gosset wrote under the pen name “Student.”

Up until the mid-1970s, some statisticians used the normal distribution approximation for large sample sizes and only used the t -distribution for sample sizes of at most 30. With technology, the practice now is to use the t -distribution whenever s is used as an estimate for σ .

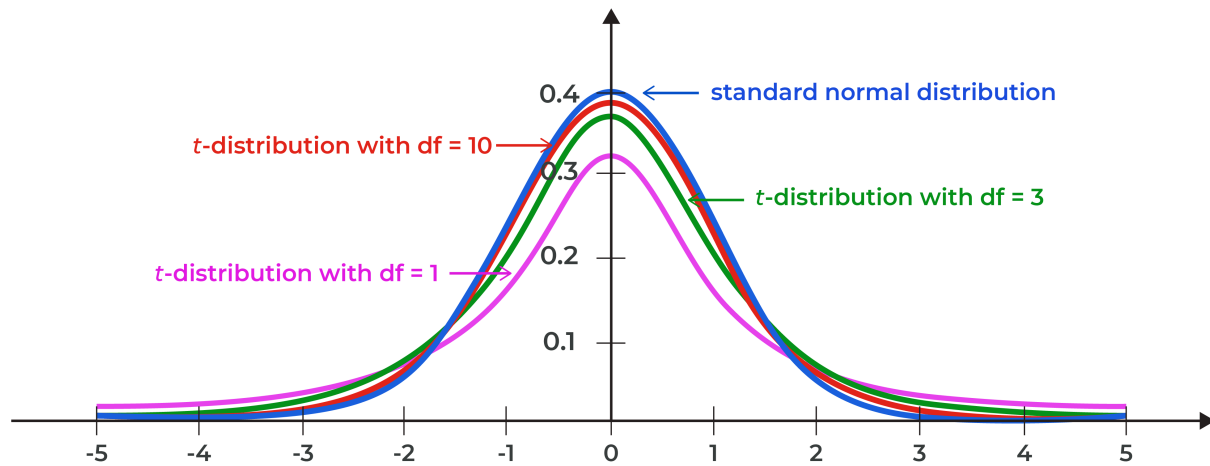
When a simple random sample of size n is taken from a population that has an approximately normal distribution with mean μ , an unknown population standard deviation, and the sample standard deviation s is used as an estimate for the population standard deviation, the distribution of the sample means follows a t -distribution with $n - 1$ degrees of freedom. For each sample size n , there is a different t -distribution. The t -score is

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Every t -distribution has a parameter called the **degrees of freedom (df)**. In this case, where the t -distribution is used for the distribution of the sample means, the value of the degrees of freedom is $n - 1$ where n is the sample size. Here the value of $n - 1$ used as the degrees of freedom comes from the calculation of the sample standard deviation s . Because the sum of the deviations is zero, we can find the last deviation once we know the other $n-1$ deviations. The other $n-1$ deviations can change or vary freely. Note that the value or formula of the degrees of freedom for the t -distribution will vary depending on the situation in which the t -distribution is used.

Properties of the t -Distribution

- The mean for the t -distribution is 0.
- The graph for the t -distribution is a symmetric, bell-shaped curve, similar to the standard normal curve. The graph is symmetric about the mean 0.
- The t -distribution has more probability in its tails than the standard normal distribution because the spread of the t -distribution is greater than the spread of the standard normal distribution. So, the graph of the t -distribution will be thicker in the tails and shorter in the centre than the graph of the standard normal distribution.
- The exact shape of the t -distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of t -distribution becomes more like the graph of the standard normal distribution. In fact, the t -distribution with an infinite number of degrees of freedom is the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with an unknown population mean μ and unknown population standard deviation σ . The size of the underlying population is generally not relevant unless it is very small. If it is bell-shaped (normal), then the assumption is met and does not need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.



Constructing the Confidence Interval

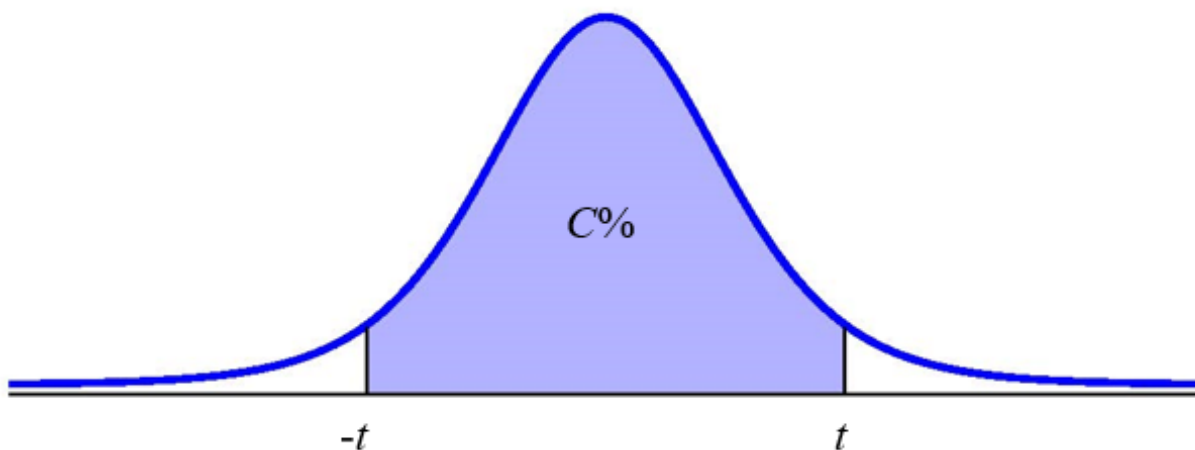
When finding a confidence interval for an unknown population mean when the population standard deviation is unknown, we use the sample standard deviation s as an estimate for the (unknown) population standard deviation, and we use a t -distribution with $n - 1$ degrees of freedom to find the required t -score for the confidence interval. In this case, we replace the z -score with a t -score and σ with s in the formulas for the limits of the confidence interval for a population mean.

To construct the confidence interval, take a random sample of size n from the population. Calculate the sample mean \bar{x} and the sample standard deviation s . The limits for the confidence interval with confidence level C for an unknown population mean μ when the population standard deviation σ is **unknown** are

$$\text{Lower Limit} = \bar{x} - t \times \frac{s}{\sqrt{n}}$$

$$\text{Upper Limit} = \bar{x} + t \times \frac{s}{\sqrt{n}}$$

where t is the (positive) t -score of the t -distribution with $n - 1$ degrees of freedom so the area under the t -distribution in between $-t$ and t is the confidence level C .



CALCULATING THE t -SCORE FOR A CONFIDENCE INTERVAL IN EXCEL

To find the t -score to construct a confidence interval with confidence level C , use the **t.inv.2t(area in the tails, degrees of freedom)** function.

- For **area in the tails**, enter the **sum** of the area in the tails of the t -distribution. For a confidence interval, the area in the tails is $1 - C$.
- For **degrees of freedom**, enter the value of the degrees of freedom for the t -distribution. For a confidence interval for a population mean, the degrees of freedom is $n - 1$.

The output from the **t.inv.2t** function is the value of the t -score needed to construct the confidence interval.

Visit the Microsoft page for more information about the **t.inv.2t** function.

NOTE

The **t.inv.2t** function requires that we enter the **sum** of the area in **both** tails. The area in the middle of the distribution is the confidence level C , so the sum of the area in both tails is the leftover area $1 - C$.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=175#oembed-1>

Video: “Confidence Interval for a population mean – t distribution” by Joshua Emmanuel [7:40] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

EXAMPLE

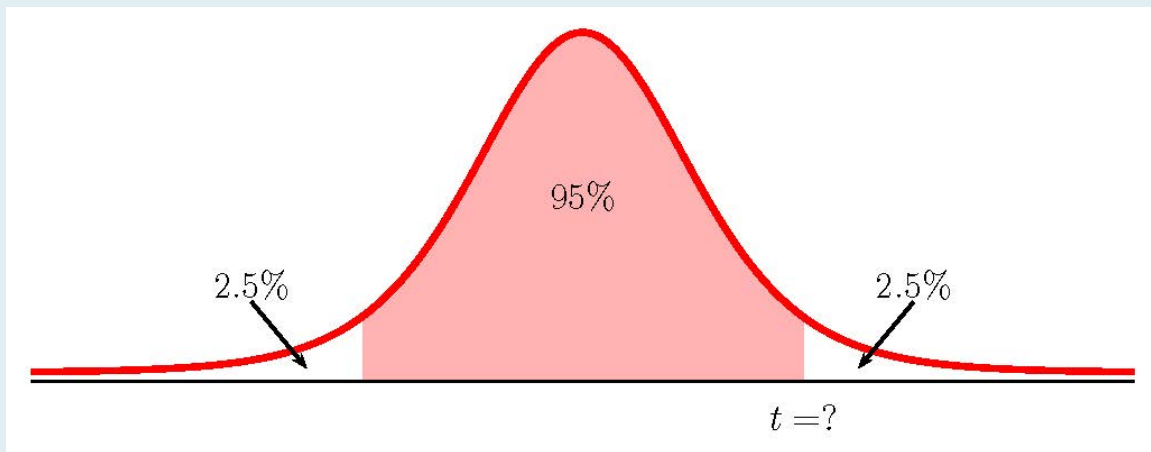
Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given below.

Sensory Rate Results		
8.6	7.3	10.3
9.4	9.2	5.4
7.9	9.6	8.1
6.8	8.7	5.5
8.3	11.4	6.9

1. Construct a 95% confidence interval for the mean sensory rate.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to conclude that the mean sensory rate is 10? Explain.

Solution

1. To find the confidence interval, we need to find the t -score for the 95% confidence interval. This means that we need to find the t -score so that the area in the tails is $1 - 0.95 = 0.05$. The degrees of freedom for the t -distribution is $n - 1 = 15 - 1 = 14$.



Function	t.inv.2t
Field 1	0.05
Field 2	14
Answer	2.1447...

So $t = 2.1447 \dots$. From the sample data supplied in the question $\bar{x} = 8.226 \dots$, $s = 1.672 \dots$ and $n = 15$. The 95% confidence interval is

$$\begin{aligned}
 \text{Lower Limit} &= \bar{x} - t \times \frac{s}{\sqrt{n}} \\
 &= 8.226\ldots - 2.1447\ldots \times \frac{1.672\ldots}{\sqrt{15}} \\
 &= 7.30
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \bar{x} + t \times \frac{s}{\sqrt{n}} \\
 &= 8.226\ldots + 2.1447\ldots \times \frac{1.672\ldots}{\sqrt{15}} \\
 &= 9.15
 \end{aligned}$$

2. We are 95\% confident that the mean sensory rate is between 7.30 and 9.15.
3. It is not reasonable to conclude that the mean sensory rate is 10 because 10 is outside of the confidence interval.

NOTE

When calculating the limits for the confidence interval, keep all of the decimals in the t -score and other values such as \bar{x} and s throughout the calculation. This will ensure that there is no round-off error in the answers. Use Excel to do the calculation of the limits, clicking on the cells containing the t -score, \bar{x} and s , to ensure that all of the decimal places are used in the calculation.

TRY IT

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results.

Hours of Sleep			
8.2	8.6	8.9	9.2
9.1	6.9	9.9	7.5
7.7	11.2	10.1	10.5

1. Construct a 97\% confidence interval for the mean number of hours slept each night.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to assume that the mean number of hours slept each night is 9 hours? Explain.

Click to see Solution

1.	Function	t.inv.2t
	Field 1	0.03
	Field 2	11
	Answer	2.4906...

$$\begin{aligned}
 \text{Lower Limit} &= \bar{x} - t \times \frac{s}{\sqrt{n}} \\
 &= 8.9833\ldots - 2.4906\ldots \times \frac{1.2904\ldots}{\sqrt{12}} \\
 &= 8.056
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \bar{x} + t \times \frac{s}{\sqrt{n}} \\
 &= 8.9833\ldots + 2.4906\ldots \times \frac{1.2904\ldots}{\sqrt{12}} \\
 &= 9.911
 \end{aligned}$$

2. We are 97\% confident that the mean number of hours slept each night is between 8.056 hours and 9.911 hours.
3. It is reasonable to assume the mean number of hours slept each night is 9 hours because 9 is inside the confidence interval.

EXAMPLE

The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the “In utero/newborn” group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. This table shows how many of the targeted chemicals were found in each infant’s cord blood.

Targeted Chemicals									
79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

- Construct a 90\% confidence interval for the mean number of targeted industrial chemicals found in an infant's blood.
- Interpret the confidence interval found in part 1.

Solution

- To find the confidence interval, we need to find the t -score for the 90\% confidence interval. This means that we need to find the t -score so that the area in the tails is $1 - 0.90 = 0.1$. The degrees of freedom for the t -distribution is $n - 1 = 20 - 1 = 19$

Function	t.inv.2t
Field 1	0.1
Field 2	19
Answer	1.7291...

So $t = 1.7291 \dots$. From the sample data supplied in the question $\bar{x} = 127.45$, $s = 25.9645 \dots$ and $n = 20$. The 90\% confidence interval is

$$\begin{aligned}
 \text{Lower Limit} &= \bar{x} - t \times \frac{s}{\sqrt{n}} \\
 &= 127.45 - 1.7291 \dots \times \frac{25.9645 \dots}{\sqrt{20}} \\
 &= 117.41
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \bar{x} + t \times \frac{s}{\sqrt{n}} \\
 &= 127.45 + 1.7291 \dots \times \frac{25.9645 \dots}{\sqrt{20}} \\
 &= 137.49
 \end{aligned}$$

- We are 90\% confident that the mean number of targeted industrial chemicals found in an

infant's blood is between 117.41 and 137.49.

TRY IT

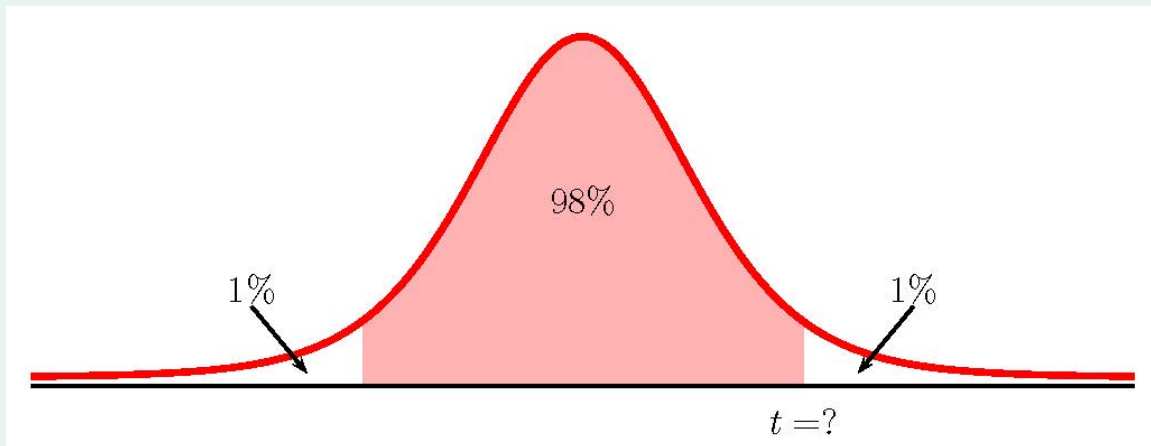
A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in this table.

Hours Watching Television				
0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

1. Construct a 98\% confidence interval for the mean number of hours statistics students will spend watching television in one week.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to conclude that the mean number of hours statistics students spend watching television in one week is 5? Explain.

Click to see Solution

1.	Function	t.inv.2t
	Field 1	0.02
	Field 2	14
	Answer	2.6244...



$$\begin{aligned}
 \text{Lower Limit} &= \bar{x} - t \times \frac{s}{\sqrt{n}} \\
 &= 6.133\dots - 2.6244\dots \times \frac{5.514\dots}{\sqrt{15}} \\
 &= 2.397
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \bar{x} + t \times \frac{s}{\sqrt{n}} \\
 &= 6.133\dots + 2.6244\dots \times \frac{5.514\dots}{\sqrt{15}} \\
 &= 9.870
 \end{aligned}$$

2. We are 98\% confident that the mean number of hours statistics students will spend watching television in one week is between **2.397** hours and **9.870** hours.
3. It is reasonable to assume the mean number of hours statistics students will spend watching television in one week is **5** hours because **5** is inside the confidence interval.

Exercises

1. A zoo keeper wants to estimate the mean weight of newborn elephants. From a sample of 50 newborn elephants, the mean weight was 122 kg with a standard deviation of 5.5 kg.

- a. Construct a 95% confidence interval for the mean weight of newborn elephants.
- b. Interpret the confidence interval found in part (a).
- c. What will happen to the confidence interval obtained if 500 newborn elephants are weighed instead of 50?

Click to see Answer

- a. Lower Limit = 120.33, Upper Limit = 123.56
- b. There is a 95% probability that the mean weight of newborn elephants is between 120.44 kg and 123.56 kg.
- c. The confidence interval will get narrower.

2. A researcher in Sweden wants to estimate the mean height of adult Swedish men. The researcher takes a sample of 45 adult Swedish men and finds the mean height in the sample is 177.5 cm with a standard deviation of 7 cm.
 - a. Construct a 98% confidence interval for the mean height of adult Swedish men.
 - b. Interpret the confidence interval found in part (a).
 - c. The research claims that the mean height of adult Swedish men is 187.5 cm. Can the research make this claim? Explain.

Click to see Answer

- a. Lower Limit = 174.98, Upper Limit = 180.02
- b. There is a 98% probability that the mean height of adult Swedish men is between 174.98 cm and 180.02 cm.
- c. No, because 187.5 cm is outside the confidence interval.

3. Announcements for 84 upcoming engineering conferences were randomly picked from a stack of IEEE Spectrum magazines. The mean length of the conferences was 3.94 days with a standard deviation of 1.28 days.
 - a. Construct a 97% confidence interval for the mean length of time of engineering conferences.
 - b. Interpret the confidence interval found in part (a).
 - c. Is it reasonable to claim that the mean length of time of engineering conferences is 3 days? Explain.

Click to see Answer

- a. Lower Limit = 3.63, Upper Limit = 4.25
- b. There is a 97% probability that the mean length of time of engineering conferences is between 3.63 days and 4.25 days.

c. No, because 3 cm is outside the confidence interval.

4. A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients and found that the sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.
- Construct a 99% confidence interval for the mean wait time in the emergency room.
 - Interpret the confidence interval found in part (a).
 - Is it reasonable to claim that the mean wait time is 2 hours? Explain.

Click to see Answer

- Lower Limit = 1.34, Upper Limit = 1.66
 - There is a 99% probability that the mean wait time in the emergency room is between 1.34 days and 1.66 days.
 - No, because 2 cm is outside the confidence interval.
5. A researcher wants to estimate the mean time adults spend watching television per month. In a sample of 108 adults, the mean time spent watching television per month was 151 hours with a standard deviation of 32 hours.
- Construct a 93% confidence interval for the mean time adults spend watching television per month.
 - Interpret the confidence interval found in part (a).
 - The researcher claims that the mean time adults spend watching television per month is 155 hours. Can the researcher make this claim? Explain.

Click to see Answer

- Lower Limit = 145.36, Upper Limit = 156.64
 - There is a 93% probability that the mean time adults spend watching television per month is between 145.36 hours and 156.64 hours.
 - No, because 3 cm is outside the confidence interval.
6. A researcher wants to estimate the mean enrollment at the country's post-secondary institutions. A random survey of enrollment numbers at 35 post-secondary institutions across the country yielded the following figures:

6,414	4,300	5,481	6,357	17,500	13,713	1,263
1,550	5,944	5,200	27,000	9,200	17,768	7,285
2,109	5,722	5,853	9,414	7,380	7,493	28,165
9,350	2,825	2,750	7,681	18,314	2,771	5,080
21,828	2,044	10,012	3,200	6,557	2,861	11,622

- Construct a 95\% confidence interval for the mean enrollment at post-secondary institutions across the country.
- Interpret the confidence interval found in part (a).
- Is it reasonable to conclude that the mean enrollment at post-secondary institutions across the country is 15,000 students? Explain.
- What will happen to the confidence interval if 500 post-secondary institutions were surveyed instead of 35?

Click to see Answer

- Lower Limit = 6,243.44, Upper Limit = 11,014.05
- There is a 95\% probability that the mean enrollment at post-secondary institutions across the country is between 6,243.44 students and 11,014.05 students.
- No, because 15,000 students is outside the confidence interval.
- The confidence interval would get narrower.

- Suppose that a committee is studying whether or not there is a waste of time in our judicial system. The committee is interested in the mean amount of time individuals spend at the courthouse waiting to be called for jury duty. The committee randomly surveyed 81 people who recently served as jurors and found that the sample mean wait time was 8 hours with a sample standard deviation of 4 hours.

- Construct a 96\% confidence interval for the mean time individuals spend waiting to be called for jury duty.
- Explain in a complete sentence what the confidence interval means.

Click to see Answer

- Lower Limit = 7.07, Upper Limit = 8.93
- There is a 96\% probability that the mean time individuals spend waiting to be called for jury duty is between 7.07 hours and 8.93 hours.

- A pharmaceutical company makes tranquilizers. Researchers in a hospital used the drug on a

random sample of patients and recorded the length of time, in hours, that the tranquillizer lasted. The data is recorded as follows:

Tranquilizer Duration (hours)					
2.9	2.3	2.4	2.6	3.2	2
3	2.5	2.5	3.2	3	2.8
2	2.9	2.6	3.2	2.5	2.6
2	2.6	2.5	2.5	2.5	2.7
3.1	2.5	2.8	2.4	2.3	2.9

- Construct a 94\% confidence interval for the mean length of time the tranquillizers last.
- What does it mean to be “94\% confident” in this problem?
- Can the researchers claim that the mean length of time the tranquillizer lasts is 2.7 hours? Explain.

Click to see Answer

- Lower Limit = 2.51, Upper Limit = 2.76
- There is a 94\% probability that the mean length of time the tranquillizer lasts is between 2.51 hours and 2.76 hours.
- Yes, because 2.7 hours is inside the confidence interval.

- Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate the mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected, and the number of unoccupied seats is noted for each of the sampled flights. In the sample, the mean number of unoccupied seats is 11.6 seats with a standard deviation is 4.1 seats.

- Construct a 92\% confidence interval for the mean number of unoccupied seats per flight.
- Interpret the confidence interval found in part (a).
- Is it reasonable for the airlines to claim that the mean number of unoccupied seats per flight is 15? Explain.

Click to see Answer

- Lower Limit = 11.12, Upper Limit = 12.08
- There is a 92\% probability that the mean number of unoccupied seats per flight is between 11.12 seats and 12.08 seats.
- No, because 15 seats is outside the confidence interval.

10. A used car dealership wants to estimate the mean cost of a used car. In a recent sample of 84 used car sales costs, the sample mean was \$6,425 with a standard deviation of \$3,156.
- Construct a 98% confidence interval for the mean cost of a used car.
 - Explain what a “98% confidence interval” means for this study.
 - Can the used car dealership claim that the mean cost of a used car is \$7,000? Explain.

Click to see Answer

- Lower Limit = 5,608.17, Upper Limit = 7,241.83
 - There is a 98% probability that the mean cost of a used car is between \$5,608.17 and \$7,241.83 seats.
 - Yes, because \$7,000 is inside the confidence interval.
11. A local bargain hunter wants to estimate the mean dollar amount off that coupons provided by local retailers provide to consumers. The bargain hunter takes a random sample of coupons and records the amount off, in dollars, offered by each coupon:

Coupon Savings (in dollars)					
3.00	0.70	0.80	2.20	1.90	1.80
0.50	2.60	1.80	2.40	0.60	2.80
0.50	2.50	2.60	2.20	2.80	1.30
2.80	1.60	2.50	2.80	2.20	1.00
1.70	1.60	1.10	1.20	1.90	2.70
2.80	1.20	1.40	2.40	2.70	1.70

- Construct a 99% confidence interval for the mean dollar amount off of coupons.
- Interpret the confidence interval found in part (a).
- Can the bargain hunter claim that the mean dollar amount off provided by coupons is \$2.00. Explain.

Click to see Answer

- Lower Limit = 1.55, Upper Limit = 2.25
 - There is a 99% probability that the mean amount off of coupons is between \$1.55 and \$2.25 seats.
 - Yes, because \$2.00 is inside the confidence interval.
-

“7.3 Confidence Intervals for a Single Population Mean with Unknown Population Standard Deviation” and “7.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

7.4 CONFIDENCE INTERVALS FOR A POPULATION PROPORTION

LEARNING OBJECTIVES

- Calculate and interpret confidence intervals for estimating a population proportion.

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40\% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95\% confidence, so, the pollsters would be 95\% confident that the true proportion of voters who favoured the candidate would be between 37\% and 43\%.

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households that own personal computers.

A confidence interval for a population proportion is based on the fact that the sample proportions follow an approximately normal distribution when both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$. Similar to confidence intervals for population means, a confidence interval for a population proportion is constructed by taking a sample of size n from the population, calculating the sample proportion \hat{p} , and then adding and subtracting the margin of error from \hat{p} to get the limits of the confidence interval.

In order to construct a confidence interval for a population proportion, we must be able to assume

the sample proportions follow a normal distribution. As we have seen previously, we can assume the sample proportions follow a normal distribution when both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$. But in this situation, the population proportion p is unknown, so we cannot check the values of $n \times p$ and $n \times (1 - p)$. Because we must take a sample and calculate the sample proportion \hat{p} , we can check the quantities $n \times \hat{p}$ and $n \times (1 - \hat{p})$. For a confidence interval for a population proportion, if both $n \times \hat{p} \geq 5$ and $n \times (1 - \hat{p}) \geq 5$, we can assume the sample proportions follow a normal distribution.

Calculating the Margin of Error

The margin of error for a confidence interval with confidence level C for an unknown population proportion p is

$$\text{Margin of Error} = z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

where z is the z -score from the standard normal distribution so that the area to the left of z is $C + \frac{1 - C}{2}$.

NOTE

In the margin of error formula, the sample proportion \hat{p} is used to estimate the unknown population proportion p . The estimated sample proportion \hat{p} is used because p is the unknown quantity we are trying to estimate with the confidence interval. The sample proportion \hat{p} is calculated from the sample taken to construct the confidence interval where

$$\hat{p} = \frac{\text{number of items in the sample with characteristic of interest}}{n}$$

Constructing the Confidence Interval

The limits for the confidence interval with confidence level C for an unknown population proportion p are

$$\text{Lower Limit} = \hat{p} - z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

$$\text{Upper Limit} = \hat{p} + z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

where z is the z -score from the standard normal distribution so that the area to the left of z is $C + \frac{1 - C}{2}$.

NOTE

The confidence interval can only be used if we can assume the sample proportions follow a normal distribution. This means we must check that $n \times \hat{p} \geq 5$ and $n \times (1 - \hat{p}) \geq 5$ before constructing the confidence interval. If one of $n \times \hat{p}$ or $n \times (1 - \hat{p})$ is less than 5, we cannot construct the confidence interval.

CALCULATING THE z -SCORE FOR A CONFIDENCE INTERVAL IN EXCEL

To find the z -score to construct a confidence interval with confidence level C , use the **norm.s.inv(area to the left of z)** function.

- For **area to the left of z**, enter the **entire** area to the left of the z -score required. For a confidence interval, the area to the left of z is $C + \frac{1 - C}{2}$.

The output from the **norm.s.inv** function is the value of z -score needed to construct the confidence interval.

NOTE

The **norm.s.inv** function requires that we enter the **entire** area to the **left** of the unknown z -score. This area includes the confidence level C (the area in the middle of the distribution) plus the remaining area in the left tail $\frac{1 - C}{2}$.

EXAMPLE

Suppose that a market research firm is hired to estimate the percentage of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes – they own cell phones.

1. Construct a 95% confidence interval for the proportion of adult residents of this city who have cell phones.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to conclude that 85% of the adult residents of this city have cell phones? Explain.

Solution

1. The sample proportion is $\hat{p} = \frac{421}{500} = 0.842$. We need to check $n \times \hat{p}$ and $n \times (1 - \hat{p})$:

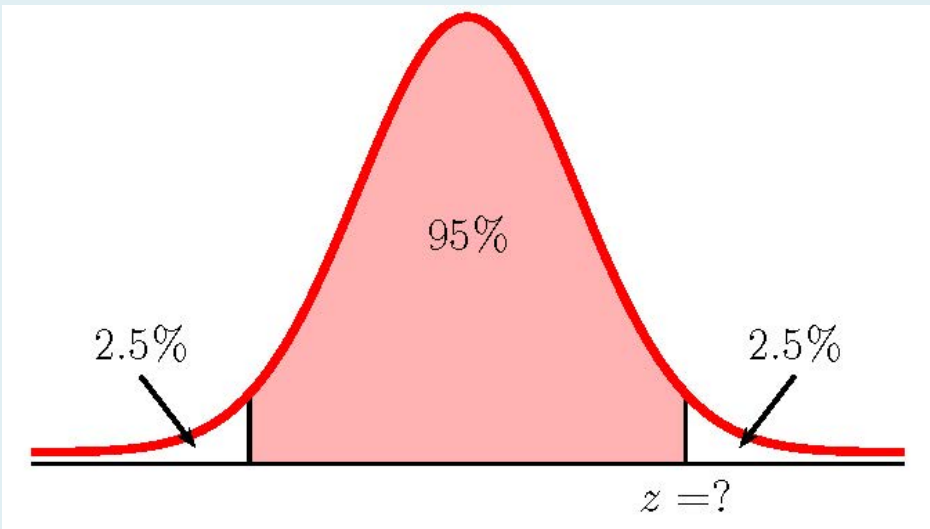
$$n \times \hat{p} = 500 \times 0.842 = 421 \geq 5$$

$$n \times (1 - \hat{p}) = 500 \times (1 - 0.842) = 79 \geq 5$$

Because both $n \times \hat{p} \geq 5$ and $n \times (1 - \hat{p}) \geq 5$, the sample proportions follow a normal distribution and we can construct the confidence interval.

To find the confidence interval, we need to find the z -score for the 95% confidence interval.

This means that we need to find the z -score from the standard normal distribution so that the entire area to the left of z is $0.95 + \frac{1 - 0.95}{2} = 0.975$.



Function	norm.s.inv
Field 1	0.975
Answer	1.9599...

So $z = 1.9599 \dots$. The 95% confidence interval is

$$\begin{aligned}
 \text{Lower Limit} &= \hat{p} - z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \\
 &= 0.842 - 1.9599... \times \sqrt{\frac{0.842 \times (1 - 0.842)}{500}} \\
 &= 0.8100
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \hat{p} + z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \\
 &= 0.842 + 1.9599... \times \sqrt{\frac{0.842 \times (1 - 0.842)}{500}} \\
 &= 0.8740
 \end{aligned}$$

2. We are 95\% confident that the proportion of adult residents of this city who have cell phones is between 81\% and 87.4\%.
3. It is reasonable to conclude that 85\% of the adult residents of this city have cell phones because 85\% is inside the confidence interval.

NOTES

1. When calculating the limits for the confidence interval, keep all of the decimals in the z -score and other values throughout the calculation. This will ensure that there is no round-off error in the answers. Use Excel to do the calculation of the limits, clicking on the cells containing the z -score and any other values, to ensure that all of the decimal places are used in the calculation.
2. The limits for the confidence interval are percents. In the above example, the upper limit of 0.8740 is the decimal form of a percent: 87.4\%.
3. When writing down the interpretation of the confidence interval, make sure to include the confidence level, and the actual population proportion captured by the confidence interval

(i.e. be specific to the context of the question), and express the limits as percents.

4. With a confidence level of $C\%$, $C\%$ of all confidence intervals constructed contain the true population parameter. In the above example, this means that 95% of all confidence intervals constructed this way contain the proportion of adult residents in this city that have a cell phone. In other words, if we constructed **100** of these confidence (using **100** different samples of size **500**), we would expect **95** of them to contain the true proportion of adult residents in this city that have a cell phone.

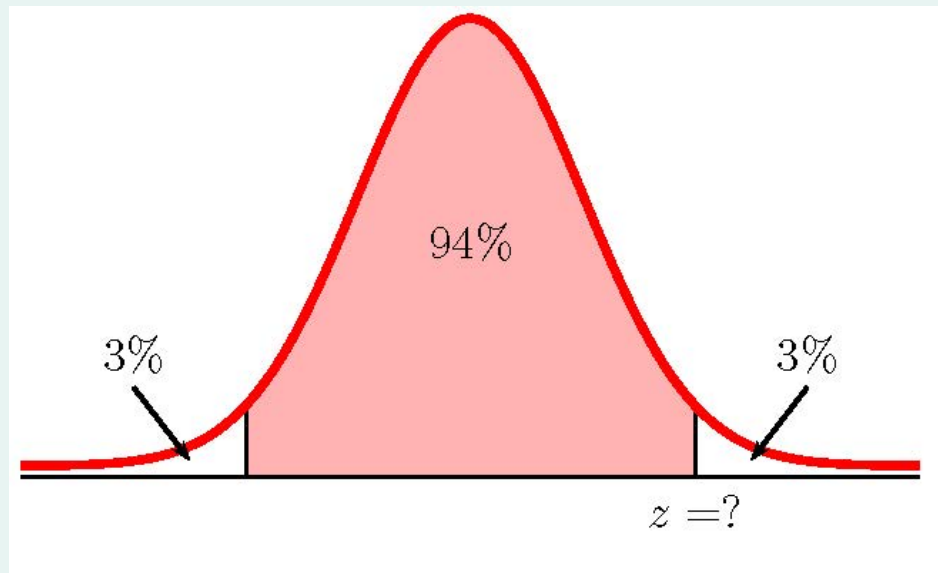
TRY IT

Suppose **250** randomly selected people are surveyed to determine if they own a tablet. Of the **250** surveyed, **98** reported owning a tablet.

1. Construct a 94% confidence interval for the proportion of people who own tablets.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to assume that 30% of people own tablets? Explain.

Click to see Solution

1.	Function	norm.s.inv
	Field 1	0.97
	Answer	1.8807...



$$\begin{aligned}
 \text{Lower Limit} &= \hat{p} - z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \\
 &= 0.392 - 1.8807... \times \sqrt{\frac{0.392 \times (1 - 0.392)}{250}} \\
 &= 0.3339
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \hat{p} + z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \\
 &= 0.392 + 1.8807... \times \sqrt{\frac{0.392 \times (1 - 0.392)}{250}} \\
 &= 0.4501
 \end{aligned}$$

2. We are 94\% confident that the proportion of people who own tablets is between 33.39\% and 45.01\%.
3. It is not reasonable to claim the proportion of people who own tablets is 30\% because 30\% is outside the confidence interval.

EXAMPLE

For a class project, a political science student at a large university wants to estimate the percentage of students who are registered voters. He surveys 500 students and finds that 300 are registered voters.

1. Construct a 90% confidence interval for the percent of students who are registered voters.
2. Interpret the confidence interval found in part 1.

Solution

1. The sample proportion is $\hat{p} = \frac{300}{500} = 0.6$. We need to check $n \times \hat{p}$ and $n \times (1 - \hat{p})$:

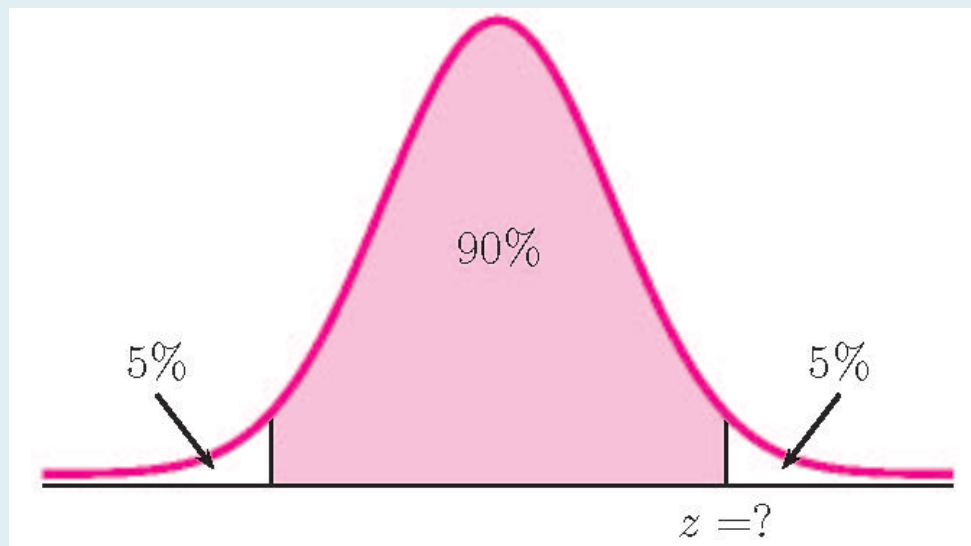
$$n \times \hat{p} = 500 \times 0.6 = 300 \geq 5$$

$$n \times (1 - \hat{p}) = 500 \times (1 - 0.6) = 200 \geq 5$$

Because both $n \times \hat{p} \geq 5$ and $n \times (1 - \hat{p}) \geq 5$, the sample proportions follow a normal distribution and we can construct the confidence interval.

To find the confidence interval, we need to find the z -score for the 90% confidence interval.

This means that we need to find the z -score from the standard normal distribution so that the entire area to the left of z is $0.90 + \frac{1 - 0.90}{2} = 0.95$.



Function	norm.s.inv
Field 1	0.95
Answer	1.6448...

So $z = 1.6448\dots$. The 90% confidence interval is

$$\begin{aligned}
 \text{Lower Limit} &= \hat{p} - z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \\
 &= 0.6 - 1.6448\dots \times \sqrt{\frac{0.6 \times (1 - 0.6)}{500}} \\
 &= 0.5640
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \hat{p} + z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \\
 &= 0.6 + 1.6448\dots \times \sqrt{\frac{0.6 \times (1 - 0.6)}{500}} \\
 &= 0.6360
 \end{aligned}$$

- We are 90% confident that the percent of students who are registered voters is between 56.4% and 63.6%.

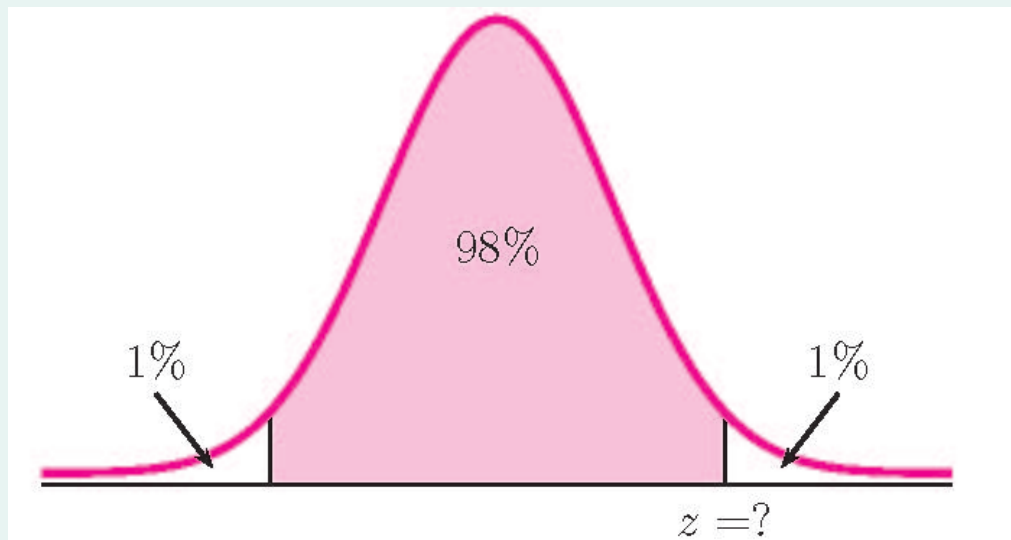
TRY IT

A student polls her school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

1. Construct a 98% confidence interval for the proportion of students who are against the new legislation.
2. Interpret the confidence interval found in part 1.
3. A parent's group claims that only 75% of students are against the legislation. Is it reasonable for the group to make this claim? Explain.

Click to see Solution

1. Function	norm.s.inv
Field 1	0.99
Answer	2.3263...



$$\begin{aligned}\text{Lower Limit} &= \hat{p} - z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \\ &= 0.8 - 2.3264 \dots \times \sqrt{\frac{0.8 \times (1 - 0.8)}{600}} \\ &= 0.7620\end{aligned}$$

$$\begin{aligned}\text{Upper Limit} &= \hat{p} + z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \\ &= 0.8 + 2.3263 \dots \times \sqrt{\frac{0.8 \times (1 - 0.8)}{600}} \\ &= 0.8380\end{aligned}$$

2. We are 98\% confident that the proportion of students who are against the new legislation is between 76.20\% and 83.80\%.
3. It is not reasonable for the group to claim the proportion is 75\% because 75\% is outside of the confidence interval.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=178#oembed-1>

Video: “Excel Statistics 85: Confidence Intervals for Proportions #1” by excelisfun [8:34] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=178#oembed-2>

Video: “Excel Statistics 86: Confidence Intervals for Proportions #2” by excelisfun [4:52] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A company manufactures individual fruit snack bags with each bag containing a mixture of grape, orange, berry, and banana-flavoured snack pieces. A sample of bags contained a total of 68 snack pieces, 5 of which were berry flavoured.
 - a. Construct a 96\% confidence interval for the proportion of berry-flavoured snack pieces per bag.
 - b. Interpret the confidence interval found in part (a).
 - c. Is it reasonable for the company to claim that each bag contains 8\% berry-flavoured snack pieces? Explain.

Click to see Answer

- a. Lower Limit = 0.0085, Upper Limit = 0.1385
 - b. There is a 96\% probability that the proportion of berry-flavoured snack pieces per bag is between 0.85\% and 13.85\%.
 - c. Yes, because 8\% is inside the confidence interval.

2. Suppose the marketing company is interested in the proportion of women who make the majority of the household purchasing decisions. They randomly surveyed 200 households and found that in 120 of them, the woman made the majority of the household purchasing decisions.
 - a. Construct a 95\% confidence interval for the proportion of households where the women make the majority of the purchasing decisions.
 - b. Interpret the confidence interval found in part (a).
 - c. Is it reasonable for the marketing company to claim that women make the majority of purchasing decisions in 70\% of households? Explain.

Click to see Answer

- a. Lower Limit = 0.5321, Upper Limit = 0.6679
 - b. There is a 95\% probability that the proportion of women who make the majority of the household purchasing decisions is between 53.21\% and 66.79\%.
 - c. No, because 70\% is outside the confidence interval.
3. A pollster is interested in what voters think is the most important issue in an upcoming election. In a poll of 1,200 voters, 65\% said the economy was the most important issue.
- a. Construct a 90\% confidence interval for the proportion of voters who believe the economy is the most important issue in the upcoming election.
 - b. Interpret the confidence interval found in part (a).
 - c. Is it reasonable to claim that 60\% of voters believe the economy is the most important issue in the upcoming election? Explain.
 - d. What would happen to the confidence interval if the level of confidence were 95\%?

Click to see Answer

- a. Lower Limit = 0.6274, Upper Limit = 0.6726
 - b. There is a 90\% probability that the proportion of voters who believe the economy is the most important issue in the upcoming election is between 62.74\% and 67.26\%.
 - c. No, because 60\% is outside the confidence interval.
 - d. The confidence interval would get wider.
4. The Ice Chalet skating school offers dozens of different ice-skating classes at various levels of ability. The Ice Chalet wants to know the proportion of girls, aged 8 to 12, in all of their ice-skating classes. In a sample of 80 skating students, 64 were girls, aged 8 to 12.
- a. Construct a 92\% confidence interval for the proportion of girls, aged 8 to 12, in the ice-skating classes.
 - b. Interpret the confidence interval found in part (a).

Click to see Answer

- a. Lower Limit = 0.7217, Upper Limit = 0.8783
 - b. There is a 92\% probability that the proportion of girls, aged 8 to 12, in the ice-skating classes is between 72.17\% and 87.83\%.
5. A university conducted a study of whether running is healthy for adults over age 50. During the eight-year study, 1.5\% of the 451 members of the 50-Plus Fitness Association died.

- a. Construct a 97% confidence interval for the proportion of adults over 50 who ran and died in the same eight-year period.
- b. Explain what a “97% confidence interval” means for this study.

Click to see Answer

- a. Lower Limit = 0.0026, Upper Limit = 0.0274
- b. There is a 97% probability that the proportion of adults over 50 who ran and died in the eight-year period is between 0.26% and 2.74%.

6. A national news magazine conducts a poll of 1,000 adults across the country. One of the questions asked in the poll was, “What is the main problem facing the country?” 20% of those polled answered “crime.”
 - a. Construct a 93% confidence interval for the proportion of adults in the country who feel that crime is the main problem.
 - b. Interpret the confidence interval found in part (a).
 - c. Is it reasonable to claim that 30% of adults feel crime is the main problem? Explain.

Click to see Answer

- a. Lower Limit = 0.1771, Upper Limit = 0.2229
- b. There is a 93% probability that the proportion of adults in the country who feel that crime is the main problem is between 17.71% and 22.29%.
- c. No, because 30% is outside the confidence interval.

7. A city’s mayor wants to estimate the proportion of residents who believe that education is one of the top issues facing the city. In a survey of 506 adult city residents, 400 said that education is a top issue for the city.
 - a. Construct a 98% confidence interval for the proportion of adult city residents who believe education is one of the top issues in the city.
 - b. Interpret the confidence interval found in part (a).
 - c. The mayor claims that 90% of adult city residents feel education is one of the top issues in the city. Is the mayor right? Explain.

Click to see Answer

- a. Lower Limit = 0.7484, Upper Limit = 0.8326
- b. There is a 98% probability that the proportion of adult city residents who believe education is one of the top issues in the city is between 74.84% and 83.26%.
- c. No, because 90% is outside the confidence interval.

8. A bank wants to study the retirement savings plans of young adults between the ages of 25 and 35. In a sample of 300 young adults between the ages of 25 and 35, only 20 made regular contributions to some type of retirement savings plan.
- Construct a 99% confidence interval for the proportion of young adults between the ages of 25 and 35 who made regular contributions to some type of retirement savings plan.
 - Interpret the confidence interval found in part (a).
 - Can the bank claim that 10% of young adults between the ages of 25 and 35 make regular contributions to some type of retirement savings plan? Explain.
 - Without performing any calculations, describe how the confidence interval would change if the confidence level changed from 99% to 90%.

Click to see Answer

- Lower Limit = 0.0296, Upper Limit = 0.1038
- There is a 99% probability that the proportion of young adults between ages 25 and 35 who make regular contributions to some type of retirement savings plan is between 2.96% and 10.38%.
- Yes, because 10% is inside the confidence interval.
- The confidence interval would get narrower.

“7.4 Confidence Intervals for a Population Proportion” and “7.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

7.5 CALCULATING THE SAMPLE SIZE FOR A CONFIDENCE INTERVAL

LEARNING OBJECTIVES

- Calculate the minimum sample size required to estimate a population parameter.

Usually, we have no control over the sample size of a data set. However, if we are able to set the sample size, as in cases where we are taking a survey, it is very helpful to know just how large the sample should be to provide the most information. Sampling can be very costly, in time, product, and money. For example, simple telephone surveys will cost approximately \$30.00 each, and some sampling requires the destruction of the product. Selecting a sample that is too large is expensive and time-consuming. But selecting a sample that is too small can lead to inaccurate conclusions. We want to find the minimum sample size required to achieve the desired level of accuracy in a confidence interval.

Calculating the Sample Size for a Population Mean

The margin of error E for a confidence interval for a population mean is

$$E = \frac{z \times \sigma}{\sqrt{n}}$$

where z is the z -score so that the area under the standard normal distribution is between $-z$ and z is the confidence level C .

Rearranging this formula, we get a formula for the sample size n :

$$n = \left(\frac{z \times \sigma}{E} \right)^2$$

In order to use this formula, we need values for z , E and σ :

- The value for z is determined by the confidence level of the confidence interval, calculated the same way we calculate the z -score for a confidence interval.
- The value for the margin of error E is set as the predetermined acceptable error, or tolerance, for the difference between the sample mean \bar{x} and the population mean μ . In other words, E is set to the maximum allowable width of the confidence interval.
- An estimate for the population standard deviation σ can be found by one of the following methods:
 - Conduct a small pilot study and use the sample standard deviation from the pilot study as an estimate for σ .
 - Use the sample standard deviation from previously collected data as an estimate for σ . Although crude, this method of estimating the standard deviation may help reduce costs significantly.
 - Use $\frac{\text{Range}}{4}$ as an estimate for σ , where **Range** is the difference between the maximum and minimum values of the population under study.

NOTES

1. Although we do not know the population standard deviation when calculating the sample size, we do not use the t -distribution in the sample size formula. In order to use the t -distribution in this situation, we need the degrees of freedom $n - 1$. But n is the sample size we are trying to estimate. So, we must use the normal distribution to determine the sample size.
2. The value of n determined from the formula is the **minimum** sample size required to achieve the desired level of confidence. The sample size n is a count, and so is an integer. It would be unusual for the value of n generated by the formula to be an integer. Because n is the minimum sample size required, we must **round** the output from the formula **up** to the next integer. If we round the value of n down, the sample size will be below the minimum required sample size.
3. After we find the sample size n and collected the data for the sample, we use the

appropriate confidence interval formula and the sample standard deviation from the actual sample (assuming σ is unknown), and not the estimate of the standard deviation used in the calculation of the sample size.

CALCULATING THE z -SCORE FOR SAMPLE SIZE IN EXCEL

To find the z -score to calculate the sample size for a confidence interval with confidence level C , use the **norm.s.inv(area to the left of z)** function.

- For **area to the left of z**, enter the **entire** area to the left of the z -score required. For a confidence interval, the area to the left of z is $C + \frac{1 - C}{2}$.

The output from the **norm.s.inv** function is the value of z -score needed to find the sample size.

NOTE

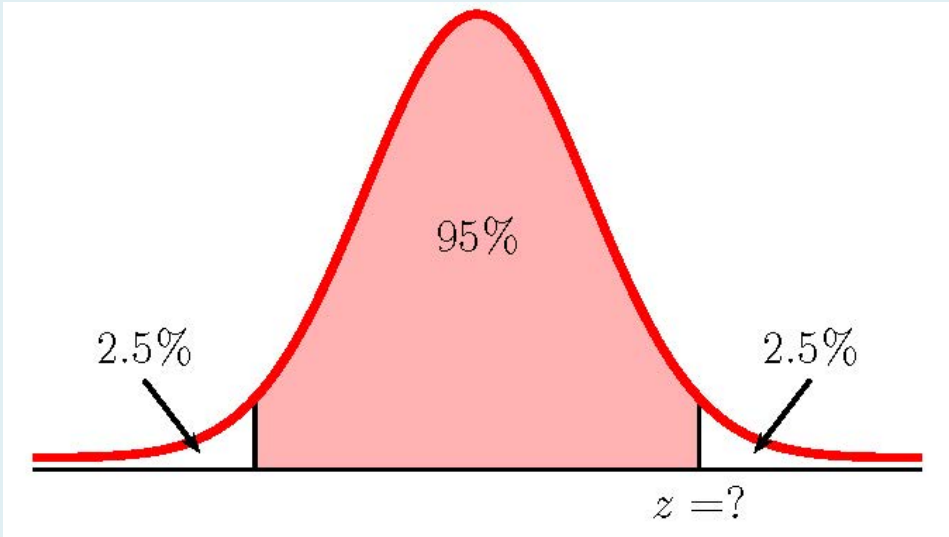
The **norm.s.inv** function requires that we enter the **entire** area to the **left** of the unknown z -score. This area includes the confidence level C (the area in the middle of the distribution) plus the remaining area in the left tail $\frac{1 - C}{2}$.

EXAMPLE

We want to estimate the mean age of Foothill College students. From previous information, an estimate of the standard deviation of the ages of the students is 15 years. We want to be 95% confident that the sample mean age is within two years of the population mean age. How many randomly selected Foothill College students must be surveyed to achieve the desired level of accuracy?

Solution

To find the sample size, we need to find the z -score for the 95% confidence interval. This means that we need to find the z -score from the standard normal distribution so that the entire area to the left of z is $0.95 + \frac{1 - 0.95}{2} = 0.975$.



Function	norm.s.inv
Field 1	0.975
Answer	1.9599...

So $z = 1.9599 \dots$. From the question $\sigma \simeq 15$ and $E = 2$.

$$\begin{aligned}
 n &= \left(\frac{z \times \sigma}{E} \right)^2 \\
 &= \left(\frac{1.9599 \dots \times 15}{2} \right)^2 \\
 &= 216.08 \dots \\
 &\Rightarrow 217 \text{ students}
 \end{aligned}$$

217 students must be surveyed to achieve the desired accuracy.

NOTE

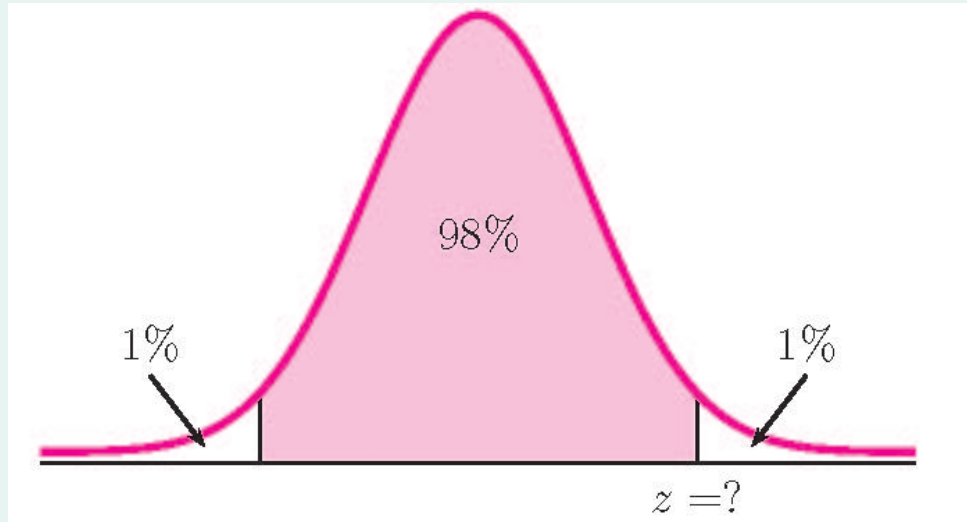
Remember to round the value for the sample size **UP** to the next integer. This ensures that the sample size is an integer and is large enough. Do not forget to include appropriate units with the sample size.

TRY IT

You want to estimate the height of all high school basketball players. You want to be 98\% confident with a margin of error of **3.75**. From a small pilot study, you estimate the standard deviation to be 7.5 cm. How large a sample do you need to take to achieve the desired level of accuracy?

Click to see Solution

Function	norm.s.inv
Field 1	0.99
Answer	2.3263...



$$\begin{aligned}
 n &= \left(\frac{z \times \sigma}{E} \right)^2 \\
 &= \left(\frac{2.3263 \dots \times 7.5}{3.75} \right)^2 \\
 &= 21.6487 \dots \\
 &\Rightarrow 22 \text{ high school basketball players}
 \end{aligned}$$

Calculating the Sample Size for a Population Proportion

The margin of error E for a confidence interval for a population proportion is

$$E = z \times \sqrt{\frac{p \times (1 - p)}{n}}$$

where z is the z -score so that the area under the standard normal distribution in between $-z$ and z is the confidence level C .

Rearranging this formula, we get a formula for the sample size n :

$$n = p \times (1 - p) \times \left(\frac{z}{E} \right)^2$$

In order to use this formula, we need values for z , E and p :

- The value for z is determined by the confidence level of the interval, calculated the same way

we calculate the z -score for a confidence interval.

- The value for the margin of error E is set as the predetermined acceptable error, or tolerance, for the difference between the sample proportion \hat{p} and the population proportion p . In other words, E is set to the maximum allowable width of the confidence interval.
- An estimate for the population proportion p . If no estimate for the population proportion is provided, we use $p = 0.5$.

NOTES

1. The value of n determined from the formula is the **minimum** sample size required to achieve the desired level of confidence. The sample size n is a count, and so is an integer. It would be unusual for the value of n generated by the formula to be an integer. Because n is the minimum sample size required, we must **round** the output from the formula **up** to the next integer. If we round the value of n down, the sample size will be below the minimum required sample size.
2. After we find the sample size n and collect the data for the sample, we use the appropriate confidence interval formula and the sample proportion from the actual sample.
3. By using 0.5 as an estimate for p in the sample size formula, we will get the largest required sample size for the confidence level and margin of error selected. This is true because of all combinations of two fractions (the values of p and $1 - p$) that add to one, and the largest multiple is when each is 0.5. Without any other information concerning the population parameter p , this is the common practice. This may result in oversampling, but certainly not under-sampling.

There is an interesting trade-off between the level of confidence and the sample size that shows up here when considering the cost of sampling. The table below shows the appropriate sample size at different levels of confidence and different margins of error, assuming $p = 0.5$. Looking at each row, we can see that for the same margin of error, a higher level of confidence requires a larger sample size. Similarly, looking at each column, we can see that for the same confidence level, a smaller margin of error requires a larger sample size.

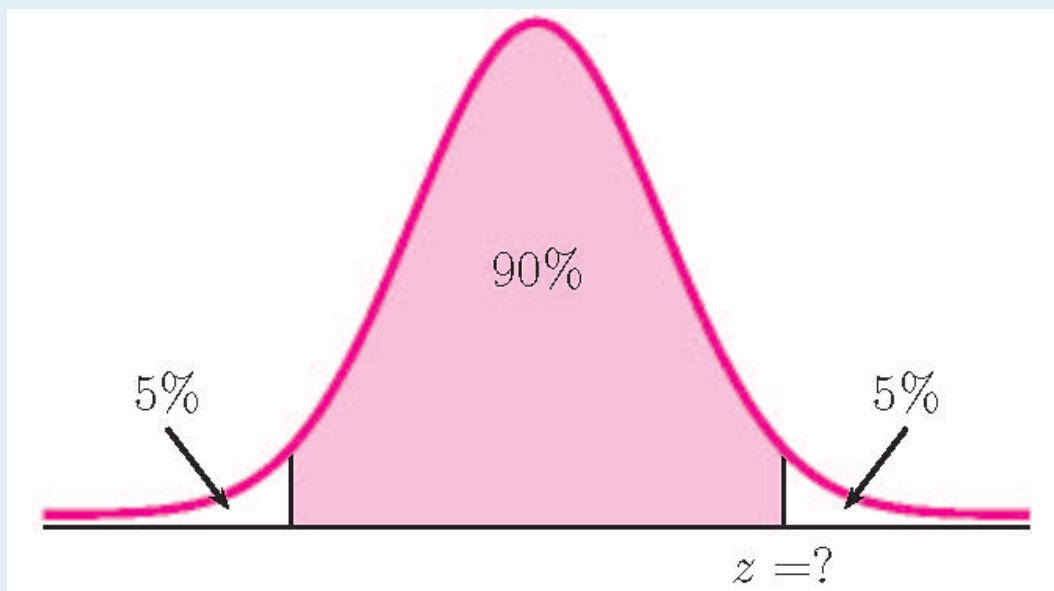
Required Sample Size (90%)	Required Sample Size (95%)	Margin of Error
1691	2401	2%
752	1067	3%
271	384	5%
68	96	10%

EXAMPLE

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident with a margin of error of 3%?

Solution

To find the sample size, we need to find the z -score for the 90% confidence interval. This means that we need to find the z -score from the standard normal distribution so that the entire area to the left of z is $0.90 + \frac{1 - 0.90}{2} = 0.95$.



Function	norm.s.inv
Field 1	0.95
Answer	1.6448...

So $z = 1.6.448. . . .$ From the question $E = 0.03$. Because no estimate of the population proportion is given, $p = 0.5$.

$$\begin{aligned}
 n &= p \times (1 - p) \times \left(\frac{z}{E} \right)^2 \\
 &= 0.5 \times (1 - 0.5) \times \left(\frac{1.6448. . .}{0.03} \right)^2 \\
 &= 751.539. . . \\
 &\Rightarrow 752 \text{ customers age } 50+
 \end{aligned}$$

752 customers aged 50+ must be surveyed to achieve the desired accuracy.

NOTE

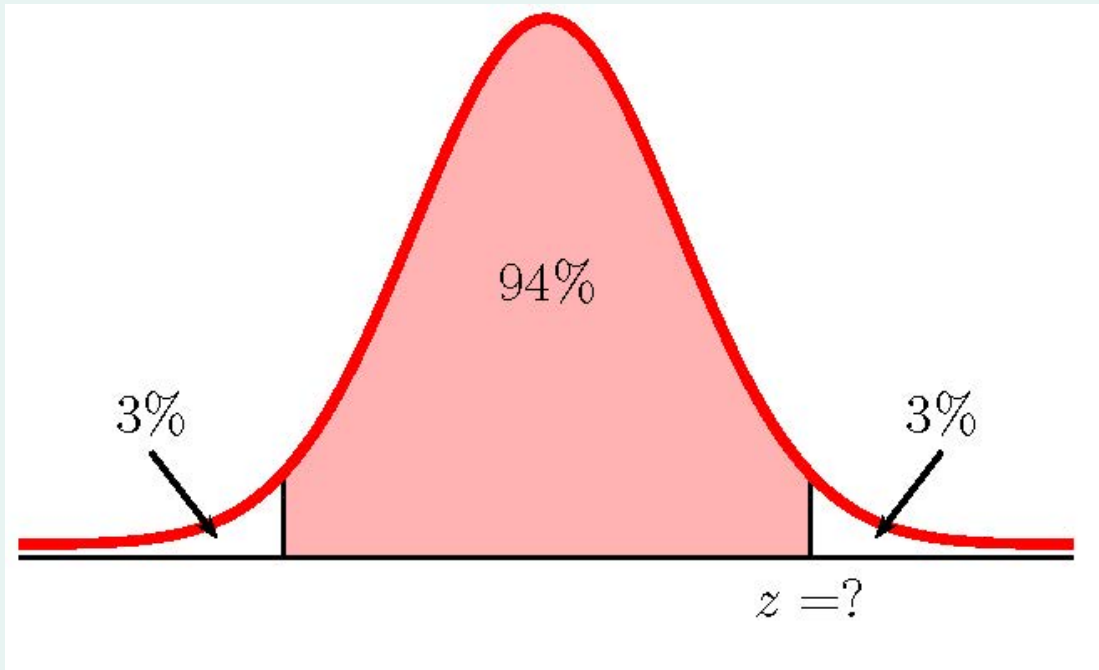
Remember to round the value for the sample size **UP** to the next integer. This ensures that the sample size is large enough. Do not forget to include appropriate units with the sample size.

TRY IT

Suppose an internet marketing company wants to determine the percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 94% confident that the estimated proportion is within 5% of the population proportion of customers who click on ads on their smartphones?

Click to see Solution

Function	norm.s.inv
Field 1	0.97
Answer	1.8807...



$$\begin{aligned}
 n &= p \times (1 - p) \times \left(\frac{z}{E} \right)^2 \\
 &= 0.5 \times (1 - 0.5) \times \left(\frac{1.8807...}{0.05} \right)^2 \\
 &= 353.738... \\
 &\Rightarrow 354 \text{ customers}
 \end{aligned}$$





One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=180#oembed-1>

Video: “Excel Statistics 87: Sample Size for Confidence Intervals” by excelisfun [7:55] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. The average height of young adult males has a normal distribution with a standard deviation of 6.25 cm. You want to estimate the mean height of students at your college or university to within 3 cm with 93% confidence. How many male students must you measure?

Click to see Answer

15 male students

2. A researcher wants to estimate the mean weight of apples in an orchard. The researcher wants to be 94% confident with a margin of error of 30 gram. The research takes a small sample of apples and estimates the standard deviation to be 120 grams. How many apples should the research sample to achieve the required accuracy?

Click to see Answer

57 apples

3. An events coordinator for a local arena wants to estimate the mean ticket price for an upcoming event. She wants to be within \$2 of the actual mean with 99% confidence. The coordinator does not have an estimate for the standard deviation of the ticket prices, but she does know that the tickets range in price from \$25 to \$75. How many tickets should the coordinator sample to achieve the required accuracy?

Click to see Answer

260 tickets

4. A marketer working for a large e-commerce company wants to estimate the mean amount of time a customer spends on the company's website. Based on a previous study, the marketer estimates the standard deviation to be approximately 4 minutes and 15 seconds. How many customers should the marketer sample if they want to be 96% confident with a margin of error of 45 seconds?

Click to see Answer

136 customers

5. A human resources manager for a very large company wants to estimate the mean length of employment for the company's employees. Currently, the HR manager knows that the length of employment for employees ranges from 2 months to 30 years. How large a sample does the HR manager need to collect in order to be within 10 months of the actual mean with 97% confidence?

Click to see Answer

434 employees

6. Insurance companies are interested in knowing the percentage of drivers who always buckle up before riding in a car.
- When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95% confident that the population proportion is estimated to be within 3%?
 - If it were later determined that it was important to be more than 95% confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?

Click to see Answer

- 752 drivers
 - The sample size would increase because a larger sample is required for a higher confidence. In this case, the minimum sample size is 1068 drivers.
7. You plan to conduct a survey on your college campus to learn about the political awareness of students. You want to estimate the true proportion of college students on your campus who voted in the last federal election with 97% confidence and a margin of error of 5%. How

many students must you interview?

Click to see Answer

471 students

8. The quality control inspector at a company that produces microchips wants to estimate the proportion of defective microchips the company produces. The inspector knows from experience that no more than 7% of the microchips produced are defective. The inspector wants a margin of error of 2%. What size sample should the inspector take in order to be 95% confident of the result?

Click to see Answer

626 microchips

9. A major airline wants to estimate the proportion of their flights that arrive late to their destination. Currently, the airline estimates that at least 20% of their flights arrive late. How large a sample does the airline need to collect to be within 6% of the actual proportion with 98% confidence?

Click to see Answer

241 flights

10. A financial planner wants to estimate the proportion of a city's residents who make regular contributions to a retirement fund. The financial planner wants to be within 8% of the actual proportion with 94% confidence. How large a sample should the planner collect?

Click to see Answer

139 residents

PART VIII

HYPOTHESIS TESTS FOR SINGLE POPULATION PARAMETERS

One of the jobs a statistician frequently performs is to make statistical inferences about populations based on samples taken from the population. The confidence intervals we learned about in the previous chapter are one way to estimate a population parameter. Another way to make a statistical inference is to make a true or false decision about a population parameter.

For example, suppose a car dealer advertises that its new small truck gets an average of 15 kilometres per litre. As a consumer, can we believe this claim? Or suppose a tutoring service claims that its method of tutoring helps 90\% of its students get an A or a B. Should parents believe this claim? What if a company says that women managers in their company earn an average of \$60,000 per year? How could we test the validity of these claims?

A statistician will make a decision, based on sound statistical analysis, about whether such claims about a population parameter are true or false. This process is called **hypothesis testing**. A hypothesis test involves collecting data from a sample, evaluating the data, and using the evidence provided by the sample data to make a decision about whether or not there is sufficient evidence to reject or not reject the null hypothesis.

Hypothesis testing consists of two contradictory hypotheses: a decision based on the data and a conclusion. To perform a hypothesis test, a statistician will:

1. Set up two contradictory hypotheses. Only one of these hypotheses is true, and the hypothesis test will determine which of the hypotheses is **most likely** true.
2. Collect sample data. (In homework problems, the data or summary statistics will be given to you.)
3. Determine the correct distribution to perform the hypothesis test.
4. Analyze the sample data by performing calculations that ultimately will allow you to reject or not reject the null hypothesis.
5. Make a decision and write a meaningful conclusion.

This chapter will focus on the hypothesis test process, how to conduct hypothesis tests on single population means and single population proportions, and errors associated with hypothesis testing.

In later chapters, we will learn how to conduct a hypothesis test on other population parameters, including population variance, two population means, two population proportions, and two population variances.

CHAPTER OUTLINE

8.1 Null and Alternative Hypotheses

8.2 The Hypothesis Test Process

8.3 Outcomes and the Type I and Type II Errors

8.4 Hypothesis Tests for a Population Mean with Known Population Standard Deviation

8.5 Hypothesis Tests for a Population Mean with Unknown Population Standard Deviation

8.6 Hypothesis Tests for a Population Proportion

“8.1 Introduction to Hypothesis Testing” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

8.1 NULL AND ALTERNATIVE HYPOTHESES

LEARNING OBJECTIVES

- Define the null and alternative hypotheses.
- State the null and alternative hypotheses in a particular situation.

A hypothesis test begins by considering **two hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints and **only one** of these hypotheses is true. The hypothesis test determines which hypothesis is **most likely** true.

- The **null hypothesis** is denoted H_0 . It is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt.
 - The null hypothesis is a claim that a population parameter equals some value. For example, $H_0 : \mu = 5$.
- The **alternative hypothesis** is denoted H_a . It is a claim about the population that is contradictory to the null hypothesis and is what we conclude is true if we reject H_0 .
 - The alternative hypothesis is a claim that a population parameter is greater than, less than, or not equal to some value. For example, $H_a : \mu > 5$, $H_a : \mu < 5$, or $H_a : \mu \neq 5$.
 - The form of the alternative hypothesis depends on the wording of the hypothesis test.
 - An alternative notation for H_a is H_1 .

Because the null and alternative hypotheses are contradictory, we must examine the evidence to decide if we have enough evidence to reject the null hypothesis or not reject the null hypothesis. In statistics, the evidence is in the form of sample data. The sample data will either support the claim that the null hypothesis is true or will be strong enough to support the claim of the alternative hypothesis. After we have determined which hypothesis the sample data supports, we make a

decision about the validity of the null hypothesis. There are two options for the **decision**. They are “**reject H_0** ” if the sample information favours the alternative hypothesis or “**do not reject H_0** ” if the sample information is insufficient to reject the null hypothesis.

EXAMPLE

A candidate in a local election claims that 30\% of registered voters voted in a recent election. Information provided by the returning office suggests that the percentage is higher than the 30\% claimed.

Solution

The parameter under study is the proportion of registered voters, so we use p in the statements of the hypotheses. The hypotheses are

$$\begin{array}{l} H_0: p = 30\% \\ H_a: p > 30\% \end{array}$$

NOTES

1. The null hypothesis H_0 is the claim that the proportion of registered voters that voted equals 30\%.
2. The alternative hypothesis H_a is the claim that the proportion of registered voters that voted is greater than (i.e. higher) than 30\%.

TRY IT

A medical researcher believes that a new medicine reduces cholesterol by 25%. A medical trial suggests that the percent reduction is different than claimed. State the null and alternative hypotheses.

Click to see Solution

$$\begin{array}{l} H_0: p = 25\% \\ H_a: p \neq 25\% \end{array}$$

EXAMPLE

We want to test whether the mean GPA of students in the nation's colleges is different from 2.0 (out of 4.0). State the null and alternative hypotheses.

Solution

$$H_0 : \mu = 2 \text{ points}$$

$$H_a : \mu \neq 2 \text{ points}$$

EXAMPLE

We want to test whether or not the mean height of eighth graders is 165 cm. State the null and alternative hypotheses.

Solution

$$H_0 : \mu = 165 \text{ cm}$$

$$H_a : \mu \neq 165 \text{ cm}$$

EXAMPLE

We want to test if college students take less than five years to graduate from college, on average. What are the null and alternative hypotheses?

Solution

$$H_0 : \mu = 5 \text{ years}$$

$$H_a : \mu < 5 \text{ years}$$

TRY IT

We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses.

Click to see Solution

$$H_0 : \mu = 45 \text{ minutes}$$

$$H_a : \mu < 45 \text{ minutes}$$

EXAMPLE

In an issue of *U.S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams. The same article stated that 6.6% of U.S. students take advanced placement exams. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

Solution

$$\begin{array}{l} H_0 : p = 6.6\% \\ H_a : p > 6.6\% \end{array}$$

TRY IT

On a state driver's test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try. State the null and alternative hypotheses.

Click to see Solution

$$\begin{array}{l} H_0: p = 40\% \\ H_a: p > 40\% \end{array}$$

Exercises

1. You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. State the null and alternative hypotheses.

Click to see Answer

$$H_0 : \mu = 3 \text{ megabits per second}$$

$$H_a : \mu > 3 \text{ megabits per second}$$

2. The mean entry-level salary of an employee at a company is \$58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.

Click to see Answer

$$H_0 : \mu = \$58,000$$

$$H_a : \mu > \$58,000$$

3. A sociologist claims the probability that a person picked at random in Times Square in New

York City is visiting the area is 83%. You want to test to see if the claim is correct. State the null and alternative hypotheses.

Click to see Answer

$$\begin{array}{l} H_0: p = 83\% \\ H_a: p \neq 83\% \end{array}$$

4. In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.

Click to see Answer

$$\begin{array}{l} H_0: p = 42\% \\ H_a: p < 42\% \end{array}$$

5. Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. If you were conducting a hypothesis test to determine if the mean length of jail time has increased, what would the null and alternative hypotheses be?

Click to see Answer

$$H_0: \mu = 2.5 \text{ years}$$

$$H_a: \mu > 2.5 \text{ years}$$

6. If you were conducting a hypothesis test to determine if the population mean time on death row could likely be 15 years, what would the null and alternative hypotheses be?

Click to see Answer

$$H_0: \mu = 15 \text{ years}$$

$$H_a: \mu > 15 \text{ years}$$

7. The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5% of American adults suffer from depression or a depressive illness. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population, what would the null and alternative hypotheses be?

Click to see Answer

$$H_0: \mu = 9.5 \quad H_a: \mu < 9.5$$

8. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. State the null and alternative hypotheses.

Click to see Answer

$H_0 : \mu = 4.5$ hours per week

$H_a : \mu > 4.5$ hours per week

“8.2 Null and Alternative Hypotheses” and “8.9 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

8.2 THE HYPOTHESIS TEST PROCESS

LEARNING OBJECTIVES

- Describe hypothesis testing in general and in practice.
- Identify the distribution required to conduct a hypothesis test.
- Define a rare event and identify how a rare event is used in a hypothesis test.
- Define the p – value and significance level and identify how they are used in determining the outcome of a hypothesis test.

Broadly speaking, a hypothesis test consists of the following steps:

1. State the null and alternative hypotheses in terms of the population parameter being tested.
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information and identify the significance level.
4. Identify the distribution required to conduct the hypothesis test. Use this distribution to calculate out the p – value.
5. Compare the p – value to the significance level and determine the outcome of the test.
6. Write down a conclusion about the outcome of the test.

Step 1: State the Null and Alternative Hypotheses

In the previous section, we looked at the null and alternative hypotheses.

- The **null hypothesis**, H_0 , is the original claim about the population parameter. The null hypothesis is a claim that a population parameter equals some value.
- The **alternative hypothesis**, H_a , is a contradictory claim about the population parameter.

The alternative hypothesis is a claim that a population parameter is greater than, less than, or not equal to some value. The form of the alternative hypothesis depends on the wording of the hypothesis test.

Recall that these hypotheses contain opposing viewpoints, and **only one** of these hypotheses is true. The purpose of the hypothesis test is to determine which hypothesis is **most likely** true.

Step 2: Determine if the Test is Left-Tail, Right-Tail, or Two-Tail

The form of the alternative hypothesis determines if the test is left-tail, right-tail or two-tail. The alternative hypothesis is one (and only one) of the following:

- The alternative hypothesis is less than ($<$) the claim from the null hypothesis. In this case, the test is a **left-tailed test**. The p – value is the area in the left tail of the corresponding distribution.
- The alternative hypothesis is greater than ($>$) the claim from the null hypothesis. In this case, the test is a **right-tailed test**. The p – value is the area in the right-tail of the corresponding distribution.
- The alternative hypothesis is not equal to (\neq) the claim from the null hypothesis. In this case, the test is a **two-tailed test**. The p – value is the sum of the area in both tails of the corresponding distribution.

Step 3: Collect the Sample Information and Identify the Significance Level

Sample information is used to determine which of the two hypotheses is **most likely** true. Collect a random sample from the population under study. In practice, the researcher collects the sample required for the test. Because collecting a sample for the purposes of learning in this book, the sample data or summary statistics are provided in the question.

The Significance Level and Rare Events

In order to determine which of the two hypotheses is true, we need to establish a **significance level**, denoted by α , **before** running the test. The significance level is the cut-off value for likely versus unlikely when compared to the p – value. But what do we mean by likely or unlikely in the context of a hypothesis test?

Suppose we make an assumption about the value of a population parameter (this assumption is the **null hypothesis**). We conduct the hypothesis under the **assumption** that the null hypothesis is true. Then, we randomly select a sample from the population. If the sample has properties that would be very **unlikely** to occur under the assumption the null hypothesis is true, then we would conclude that our assumption about the population is probably **incorrect**. Remember that our assumption is just an **assumption**—it is not a fact, and it may or may not be true. But the sample data we collect is real, and the information from that sample is a fact that may or may not support the assumption we make about the null hypothesis.

For example, Didi and Ali are at the birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside of because they will be blindfolded. There are 200 plastic bubbles in the basket, and Didi and Ali have been told that there is only one with a \$100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a \$100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is such an **unlikely** occurrence, Ali is hoping that what the two of them were told is wrong and there are more \$100 bills in the basket. In this case, a “**rare event**” has occurred (Didi getting the \$100 bill), so Ali doubts the original assumption about only one \$100 bill being in the basket.

A **rare event** is something we consider to be **unlikely** to happen (i.e. the probability of that event happening is very small). This is what we are looking for in a hypothesis test. We want to determine if the sample collected for the test is a rare event (unlikely to happen) under the assumption the null hypothesis is true. To determine if the sample is a rare event, we calculate the probability (the p – value) of the sample occurring, assuming that the null hypothesis is true. This is where the significance level comes in. We use the significance level as the “cut-off” mark for this probability. If the probability of the sample occurring is small (less than the significance level), then the sample is a “rare event” and unlikely to occur under the assumption the null hypothesis is true. In such a case, we would conclude that the original assumption that the null hypothesis is true must be incorrect, and so we would reject the null hypothesis in favour of the alternative hypothesis. If the probability of the sample occurring is not small (greater than the significance level), then the sample is not a “rare event” and is actually likely to occur under the assumption the null hypothesis is true. In this case we would conclude the original assumption that the null hypothesis is true must be correct, and so we would not reject the null hypothesis.

Remember, a rare event is an event that is unlikely to happen. But unlikely does not mean impossible. The probability of a rare event is very small (less than the significance level), which means that the chance of it happening is very small. But as long as the probability is not zero, there is still a possibility the event could happen.

In general, the significance level is a small probability. Typical significance levels used in hypothesis testing are 5\% and 1\%. As noted above, the significance level is used as the “cut-off” for likely or unlikely under the assumption the null hypothesis is true. Another way to think of the significance level is that the significance level is the probability of rejecting a null hypothesis when the null hypothesis is actually true. Rejecting a null hypothesis when it is true is called a Type I error, which is discussed in more detail in the next section.

NOTE

It is very important that the significance level is set **before** collecting and analyzing the sample, calculating the p – value, and running the test. Setting the significance level after the fact allows us to manipulate the test to get the outcome we want, which would invalidate the result.

Step 4: Identify the Distribution and Calculate the p – value

In this chapter, we will look at the hypothesis test for two different population parameters: mean and proportion. The parameter we are testing helps us determine which distribution to use in the calculation of the p – value.

Distribution for a Hypothesis Test on a Population Mean

If the hypothesis test is on a population mean, we use the distribution of the sample means in the hypothesis test. As we learned previously, the distribution of the sample means follows a normal distribution if the population the sample is taken from is normal or if the sample size is large enough ($n \geq 30$). For a hypothesis test on a population mean we use a normal distribution when the population standard deviation is known or a t -distribution when the population standard deviation is unknown.

When we perform a **hypothesis test of a single population mean** μ and the population standard deviation is **known**, we take a simple random sample from the population. We use a normal distribution, assuming the population is normal or the sample size is large enough ($n \geq 30$). The

z -score we need is the z -score from the distribution of the sample means: $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

When we perform a **hypothesis test of a single population mean** μ and the population standard deviation is **unknown**, we take a simple random sample from the population. We use a t -distribution, assuming the population is normal or the sample size is large enough ($n \geq 30$). We use the sample standard deviation to approximate the population standard deviation. The t -score we need is: $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$.

Distribution for a Hypothesis Test on a Population Proportion

If the hypothesis test is on a population proportion, we use the distribution of the sample proportions in the hypothesis test. As we learned previously, the distribution of the sample proportions follows a normal distribution if $n \times p \geq 5$ and $n \times (1 - p) \geq 5$ or a binomial distribution if one of $n \times p < 5$ or $n \times (1 - p) < 5$. For a hypothesis test on a population proportion, we use either a normal distribution or a binomial distribution, depending on which of the above conditions is met.

When we perform a **hypothesis test of a single population proportion** p , we take a simple random sample from the population. We use a normal distribution when $n \times p \geq 5$ and $n \times (1 - p) \geq 5$. In this case, the z -score we need is the z -score from the distribution of the sample proportions: $z = \sqrt{\frac{p \times (1 - p)}{n}}$. Otherwise, we use a binomial distribution when one of $n \times p < 5$ or $n \times (1 - p) < 5$.

Calculating the p - value

Once we know which distribution to use, we calculate the p - value for the test. The p - value is the actual probability of getting the selected sample under the assumption the null hypothesis is true. The p - value is the **probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample**.

The p - value is the area in the corresponding tail of the distribution based on the form of the alternative hypothesis:

- If the alternative hypothesis is less than ($<$) the claim from the null hypothesis, the p - value is the area in the left-tail of the corresponding distribution.
- If the alternative hypothesis is greater than ($>$) the claim from the null hypothesis, the p - value is the area in the right-tail of the corresponding distribution.

- If the alternative hypothesis is not equal to (\neq) the claim from the null hypothesis, the p – value is the sum of the area in both tails of the corresponding distribution.

We calculate out the p – value using the techniques learned in previous chapters about calculating probabilities in the normal distribution, t -distribution, or binomial distribution. Once we have calculated the p – value, we use the p – value to determine the outcome of the test. A large p – value (greater than the significance level) calculated from the sample data indicates that we should **not reject** the **null hypothesis**. A small p – value (less than the significance level) suggests strong evidence against the null hypothesis, and we would **reject** the null hypothesis if the evidence is strongly against it.

EXAMPLE

The customers of a local bakery claim that the height of the bakery's bread is, on average, 15 cm. The baker believes his customers are wrong and that the average height of the bread is more than 15 cm. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the **standard deviation** for the height is 1 cm, and the distribution of the heights is normal. Based on this sample, who is right: the customers or the baker?

Solution

Here, the population under study is the height of the loaves of bread, and μ is the average height of the loaves of bread.

The customers' claim is the null hypothesis: $\mu = 15$. The alternative hypothesis is the baker's claim: $\mu > 15$. In mathematical notation, the hypothesis are:

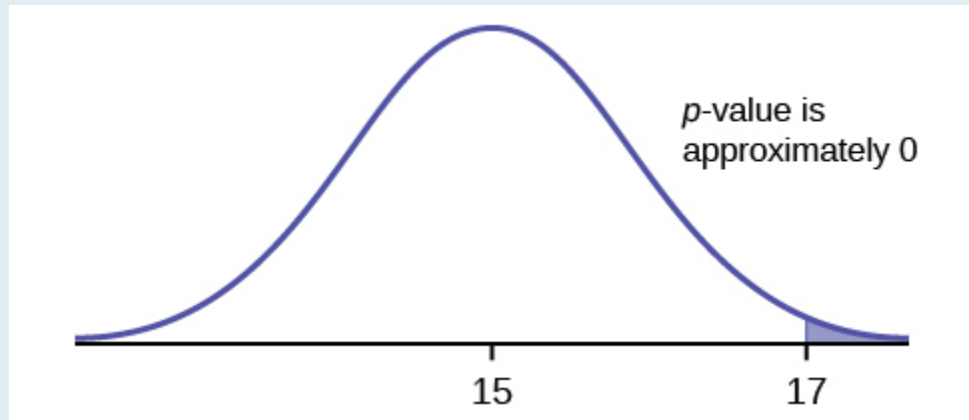
$$H_0 : \mu = 15 \text{ cm}$$

$$H_a : \mu > 15 \text{ cm}$$

Because the population standard deviation is known ($\sigma = 1$), the distribution we would use is the normal distribution.

Suppose the null hypothesis is true. That is, suppose $\mu = 15$. Under this assumption, we have to ask if the sample mean of 17 is likely or unlikely to occur. The hypothesis test works by asking the question of how **unlikely** is this sample mean if the null hypothesis is true. The graph shows how

far out the sample mean is on the normal curve. The p – value is the probability that, if we took another sample of size 10, any other sample mean would fall at least as far out as 17 cm.



The p – value is the probability that a sample mean is the same or greater than 17 cm when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for sample means. In fact, we are calculating the probability that in a sample of size 10, the sample mean is greater than 17. We learned how to calculate this type of probability when we learned about the sampling distribution of the sample mean:

Function	1-norm.dist
Field 1	17
Field 2	15
Field 3	1/sqrt(10)
Field 4	true
Answer	0.0000000001

So p – value = 0.0000000001, which tells us the probability of selecting a sample of size 10 and getting a sample mean greater than 17 is 0.0000000001 under the assumption that the null hypothesis is true ($\mu = 15$). This is a very, very small probability, which tells us that a sample mean of 17 is **unlikely** to happen if the population mean is 15 cm. Because the sample mean of 17 is so unlikely (meaning it is not happening by chance alone), we conclude that the assumption that the mean is 15 cm is wrong. That is, the evidence provided by the sample is **strongly against** the claim of the null hypothesis. So, we reject the null hypothesis in favour of the alternative hypothesis. That is, based on the test, we believe the null hypothesis is false, and the alternative hypothesis is true. So there is enough evidence to suggest that the average height of the loaves of bread is greater than 15 cm.

TRY IT

A normal distribution has a standard deviation of 1. The original claim is that the mean of the distribution is 12. An alternative claim is that the mean is greater than 12. The hypotheses are:

$$H_0 : \mu = 12$$

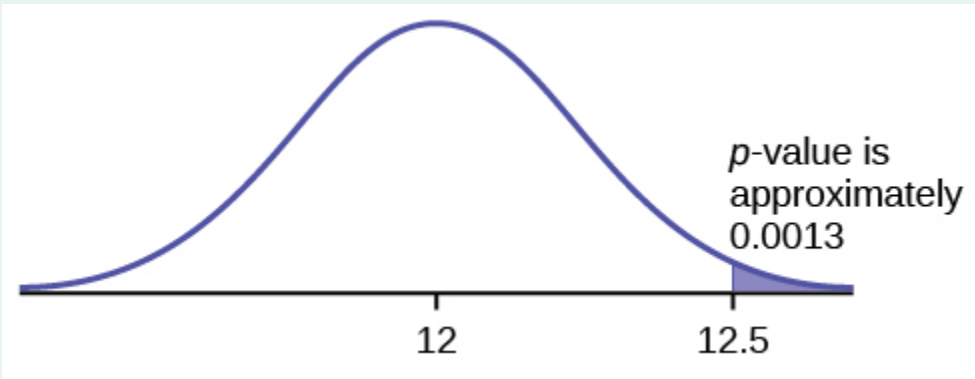
$$H_a : \mu > 12$$

In a sample of 36, the sample mean is 12.5. Calculate the p – value.

Click to see Solution

Function	1-norm.dist
Field 1	12.5
Field 2	12
Field 3	1/sqrt(36)
Field 4	true
Answer	0.0013

$$p - \text{value} = 0.0013$$



Step 5: The Outcome of the Test

The outcome of the test about whether to reject or not reject the null hypothesis is based on comparing the p – value to the significance level α . Recall that the significance level is the cut-off value for likely versus unlikely when compared to the p – value. When the p – value is greater than the significance level, the sample is likely to occur under the assumption the null hypothesis is true, and so we would fail to reject the null hypothesis. When the p – value is less than or equal to the significance level, the sample is unlikely to occur under the assumption the null hypothesis is true, and so we would reject the null hypothesis in favour of the alternative hypothesis.

When we make a **decision** to reject or not reject H_0 , do as follows:

- If p – value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
- If p – value $> \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.

When we “do not reject H_0 ,” it does not mean that we should believe that H_0 is true. It simply means that the sample data have **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of H_0 .

Step 6: Conclusion

After comparing the p – value and significance level to determine which hypothesis is most likely true, write a thoughtful **conclusion** about the hypotheses in terms of the given problem.

TRY IT

A genetics lab claims its product can increase the likelihood a pregnancy will result in a boy being born. Statisticians want to test this claim. Suppose the hypotheses are

$$\begin{array}{l} H_0: p = 50\% \\ H_a: p > 50\% \end{array}$$

After conducting the hypothesis test, $p\text{-value} = 0.025$. If the significance level is 1%, what is the conclusion of the test?

Click to see Solution

Because the $p\text{-value}$ is greater than the significance level ($p\text{-value} = 0.025 > 0.01 = \alpha$), we do not reject the null hypothesis. There is not enough evidence to support the lab's stated claim that their procedures improve the chances of a boy being born.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=192#oembed-1>

Video: “Simple hypothesis testing | Probability and Statistics | Khan Academy” by Khan Academy [6:25] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. Which distributions can you use for hypothesis testing for this chapter?

Click to see Answer

normal, t , binomial

2. Which distribution do you use when you are testing a population mean and the standard deviation is known? Assume the sample size is large.

Click to see Answer

normal

3. Which distribution do you use when the standard deviation is not known, and you are testing one population mean? Assume the sample size is large.

Click to see Answer

t

4. A population mean is 13. The sample mean is 12.8, and the sample standard deviation is 2. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.

Click to see Answer

t

5. A population has a mean is 25 and a standard deviation of 5. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?

Click to see Answer

normal

6. It is thought that 42\% of respondents in a taste test would prefer Brand A. In a particular test of 100 people, 39\% preferred Brand A. What distribution should you use to perform a

hypothesis test?

Click to see Answer

normal

7. You are performing a hypothesis test of a single population proportion. What must be true about the quantities of $n \times p$ and $n \times (1 - p)$ in order to use the normal distribution?

Click to see Answer

Both quantities must be greater than or equal to 5.

8. You are performing a hypothesis test of a single population proportion. You find out that $n \times p$ is less than five. What must you do to be able to perform a valid hypothesis test?

Click to see Answer

Use the binomial distribution.

9. When do you reject the null hypothesis?

Click to see Answer

When the p – value is less than or equal to the significance level.

10. The probability of winning the grand prize at a particular carnival game is 0.5%. Is the outcome of winning very likely or very unlikely?

Click to see Answer

very unlikely

11. The probability of winning the grand prize at a particular carnival game is 0.5%. Michele wins the grand prize. Is this considered a rare or common event? Why?

Click to see Answer

Rare because the probability of the event happening is very small.

12. What should you do when $\alpha > p - \text{value}$?

Click to see Answer

Reject the null hypothesis.

13. What should you do if $\alpha = p - \text{value}$?

Click to see Answer

Reject the null hypothesis.

14. If you do not reject the null hypothesis, then it must be true. Is this statement correct? State why or why not in complete sentences.

Click to see Answer

Incorrect because it is possible the test does not reject a false null hypothesis.

15. Assume $H_0 : \mu = 9$ and $H_a : \mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

Click to see Answer

left-tailed

16. Assume $H_0 : \mu = 6$ and $H_a : \mu > 6$. Is this a left-tailed, right-tailed, or two-tailed test?

Click to see Answer

right-tailed

17. Assume $H_0 : p = 25\%$ and $H_a : p \neq 25\%$. Is this a left-tailed, right-tailed, or two-tailed test?

Click to see Answer

two-tailed

18. Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?

Click to see Answer

two-tailed

“8.4 Distributions Required for a Hypothesis Test“, “8.5 Rare Events, The Sample, Decision, and Conclusion” and “8.9 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

8.3 OUTCOMES AND THE TYPE I AND TYPE II ERRORS

LEARNING OBJECTIVES

- Differentiate between Type I and Type II errors in a hypothesis test.

When we perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not the null hypothesis. Ideally, the hypothesis test should tell us to not reject the null hypothesis when the null hypothesis is true and reject the null hypothesis when the null hypothesis is false. However, the outcome of the hypothesis test is based on sample information and probabilities, so there is a chance that the hypothesis test does not correctly identify the truth or falseness of the null hypothesis. The outcomes are summarized in the following table:

Outcome of Test	Actual Truth State of the Null Hypothesis	
	H_0 is True	H_0 is False
Do not reject H_0	Correct Outcome	Type II Error
Reject H_0	Type I Error	Correct Outcome

The four possible outcomes in the table are:

- The decision is to **not reject** H_0 when H_0 is true (**correct decision**). That is, the test identifies H_0 is true, and in reality, H_0 is true, which means the test correctly identified H_0 as true.
- The decision is to **reject** H_0 when H_0 is true (incorrect decision known as a **Type I error**). That is, the test identifies H_0 as false, but in reality, H_0 is true, which means the test did not

correctly identify H_0 as true.

- The decision is to **not reject** H_0 when H_0 **is false** (incorrect decision known as a **Type II error**). That is, the test identifies H_0 is true, but in reality, H_0 is false, which means the test did not correctly identify H_0 as false.
- The decision is to **reject** H_0 when H_0 **is false** (**correct decision** whose probability is called the **Power of the Test**). That is, the test identifies H_0 is false, and in reality H_0 is false, which means the test correctly identified H_0 as false.

There are two types of error that can occur in hypothesis testing. Each of the errors occurs with a particular probability.

- A **Type I error** occurs when the null hypothesis is rejected by the test (i.e. the test identifies the null hypothesis as false), but in reality, the null hypothesis is true. The probability of a Type I error is the significance level α .
- A **Type II error** occurs when the null hypothesis is not rejected by the test (i.e. the test identifies the null hypothesis as true), but in reality, the null hypothesis is false. The probability of a Type II error is denoted by β .

Although the probabilities of a Type I or Type II error should be as small as possible because they are probabilities of errors, they are rarely zero.

EXAMPLE

Suppose the null hypothesis is

H_0 : Frank's rock climbing equipment is safe.

- **Type I error:** Frank thinks his rock climbing equipment is not safe when, in fact, the equipment is safe.
 - Frank believes H_0 is false, but H_0 is actually true.
- **Type II error:** Frank thinks his rock climbing equipment is safe when, in fact, the equipment is not safe.

- Frank believes H_0 is true, but H_0 is actually false.

Note that, in this case, the error with the greater consequence is the Type II error. If Frank thinks his rock climbing equipment is safe and it actually is not safe, he will go ahead and use it.

TRY IT

Suppose the null hypothesis is

H_0 : The blood cultures contain no traces of pathogen X .

State the Type I and Type II errors.

Click to see Solution

- **Type I error:** The researcher thinks the blood cultures do contain traces of pathogen X , when, in fact, they do not.
- **Type II error:** The researcher thinks the blood cultures do not contain traces of pathogen X , when in fact, they do.

EXAMPLE

Suppose the null hypothesis is

H_0 : The victim of a car accident is alive when they arrive at the ER.

- **Type I error:** The ER staff thinks that the victim is dead when, in fact, the victim is alive.
- **Type II error:** The ER staff think the victim is alive when, in fact, the victim is dead.

Note that, in this case, the error with the greater consequence is the Type I error. If the ER staff think the victim is dead, then they will not treat them.

TRY IT

Suppose the null hypothesis is

$$H_0 : \text{A patient is not sick.}$$

Which type of error has the greater consequence, Type I or Type II? Why?

Click to see Solution

The error with the greater consequence is the Type II error: the patient will be thought well when, in fact, they are sick, and so they will not get treatment.

EXAMPLE

A genetics lab claims its product can increase the likelihood a pregnancy will result in a boy being born. Statisticians want to test this claim. Suppose that the null hypothesis is

$$H_0 : \text{The genetics lab product has no effect on gender outcome.}$$

- **Type I error:** We believe the genetics lab's product can influence gender outcome when, in fact, the product has no effect.
- **Type II error:** We believe the genetics lab's product cannot influence gender outcome when, in fact, the product does have an effect.

Note that, in this case, the error with the greater consequence is the Type I error because couples would use the product in hopes of increasing the chances of having a boy.

TRY IT

“Red tide” is a bloom of poison-producing algae—a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regularly sampling shellfish along the coastline. If the mean level of toxin in clams exceeds $800\text{ }\mu\text{g}$ (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a Type I and a Type II error in this context and state which error has the greater consequence.

Click to see Solution

In this scenario, an appropriate null hypothesis would be

H_0 : The mean level of toxins is at most $800\text{ }\mu\text{g}$.

- **Type I error:** The DMF believes that toxin levels are still too high when, in fact, toxin levels are at most $800\text{ }\mu\text{g}$. The DMF continues the harvesting ban.
- **Type II error:** The DMF believes that toxin levels are within acceptable levels (are at most $800\text{ }\mu\text{g}$) when, in fact, toxin levels are still too high (more than $800\text{ }\mu\text{g}$). The DMF lifts the harvesting ban. This error could be the most serious. If the ban is lifted and clams are still

toxic, consumers could possibly eat tainted food.

In summary, the more dangerous error would be to commit a Type II error because this error involves the availability of tainted clams for consumption.

EXAMPLE

A certain experimental drug claims a cure rate of at least 75% for males with prostate cancer. Describe both the Type I and Type II errors in context. Which error is more serious?

- **Type I:** A cancer patient believes the cure rate for the drug is less than 75% when it actually is at least 75%.
- **Type II:** A cancer patient believes the experimental drug has at least a 75% cure rate when it has a cure rate that is less than 75%.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75% of the time, this will most likely influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=190#oembed-1>

Video: “Type 1 errors | Inferential statistics | Probability and Statistics | Khan Academy” by Khan Academy [3:24] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. The mean price of mid-sized cars in a region is \$32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.

Click to see Answer

- Type I Error: The mean price of mid-sized cars is believed to not be \$32,000 when in fact, the mean price actually is \$32,000.
- Type II Error: The mean price of mid-sized cars is believed to be \$32,000 when in fact, the mean price is not \$32,000.

2. A sleeping bag is tested to withstand temperatures of -30°C . You think the bag cannot withstand temperatures that low. State the Type I and Type II errors in complete sentences.

Click to see Answer

- Type I Error: The sleeping bag is believed to not withstand temperatures of -30°C when in fact, it does.
- Type II Error: The sleeping bag is believed to be able to withstand temperatures of -30°C when, in fact, it cannot.

3. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis is the surgical procedure will go well.

- a. State the Type I and Type II errors in complete sentences.
- b. Which error has the greater consequence?

Click to see Answer

- Type I Error: The doctors believe the surgical procedure will not go well when, in fact, it will.
- Type II Error: The doctors believe the surgical procedure will go well when, in fact, it will not.

- b. Type I because if the doctors believe the surgical procedure will not go well, they will not perform the operation and the patient will not get the necessary treatment.

4. A group of divers is exploring an old sunken ship. Suppose the null hypothesis is the sunken

ship does not contain buried treasure. State the Type I and Type II errors in complete sentences.

Click to see Answer

- Type I Error: The divers believe the ship contains buried treasure when, in fact, it does not.
- Type II Error: The divers believed the ship does not contain buried treasure when, in fact, it does.

5. A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis is: the sample contains E-coli. Which is the error with the greater consequence?

Click to see Answer

Type I because if the microbiologist believes the sample does not contain E-coli when in fact, it does, the public will not be advised to stop drinking the water.

6. When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the government authorities to market the drug. Suppose the null hypothesis is “the drug is unsafe.” What is the Type II error?

Click to see Answer

The drug is believed to be unsafe when, in fact, the drug is safe.

7. A statistics instructor believes that fewer than 20\% of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. What is the Type I error?

Click to see Answer

The instructor believes more than 20\% of students attended the midnight showing when in fact, less than 20\% did.

8. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The

sample mean was 4.75 hours with a sample standard deviation of 2.0. What is the Type I error?

Click to see Answer

The organization believes that the mean amount of time teenagers spend on the phone is greater than 4.5 hours per week when, in fact, the average is 4.5 hours per week.

“8.3 Outcomes and the Type I and Type II Errors” and “8.9 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

8.4 HYPOTHESIS TESTS FOR A POPULATION MEAN WITH KNOWN POPULATION STANDARD DEVIATION

LEARNING OBJECTIVES

- Conduct and interpret hypothesis tests for a population mean with known population standard deviation.

Some notes about conducting a hypothesis test:

- The null hypothesis H_0 is always an “equal to.” The null hypothesis is the original claim about the population parameter.
- The alternative hypothesis H_a is a “less than,” “greater than,” or “not equal to.” The form of the alternative hypothesis depends on the context of the question.
- The form of the alternative hypothesis tells us if the test is left-tail, right-tail, or two-tail. The alternative hypothesis is the key to conducting the test and finding the correct p – value.
 - If the alternative hypothesis is a “less than”, then the test is left-tail. The p – value is the area in the left-tail of the distribution.
 - If the alternative hypothesis is a “greater than”, then the test is right-tail. The p – value is the area in the right-tail of the distribution.
 - If the alternative hypothesis is a “not equal to”, then the test is two-tail. The p – value is the sum of the area in the two-tails of the distribution. Each tail represents exactly half of the p – value.
- **Think about the meaning of the p – value.** A data analyst (and anyone else) should have more confidence that they made the correct decision to reject the null hypothesis with a smaller p – value (for example, 0.001 as opposed to 0.04) even if using a significance level

of 0.05. Similarly, for a large p – value such as 0.4, as opposed to a p – value of 0.056 (a significance level of 0.05 is less than either number), a data analyst should have more confidence that they made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

- The significance level must be identified before collecting the sample data and conducting the test. Generally, the significance level will be included in the question. If no significance level is given, a common standard is to use a significance level of 5\%.
- An alternative approach for hypothesis testing is to use what is called the **critical value approach**. In this book, we will only use the p – value approach. Some of the videos below may mention the critical value approach, but this approach will not be used in this book.

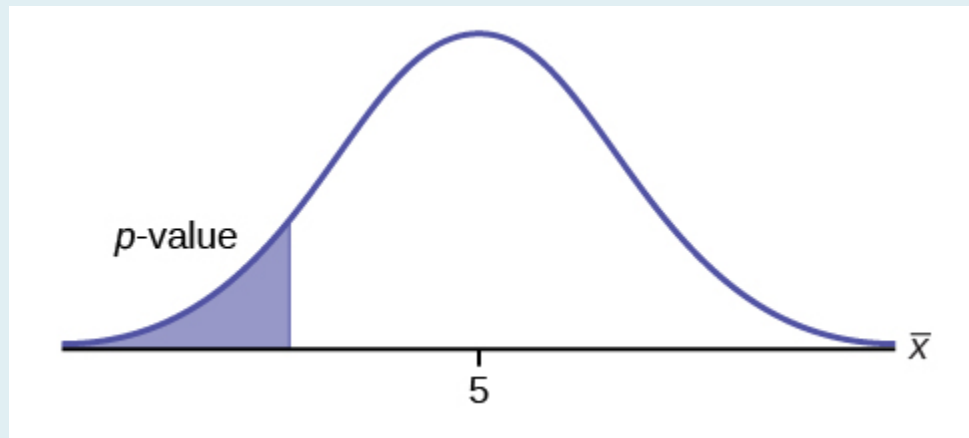
EXAMPLE

Suppose the hypotheses for a hypothesis test are:

$$H_0 : \mu = 5$$

$$H_a : \mu < 5$$

Because the alternative hypothesis is a $<$, this is a left-tailed test. The p – value is the area in the left-tail of the distribution.



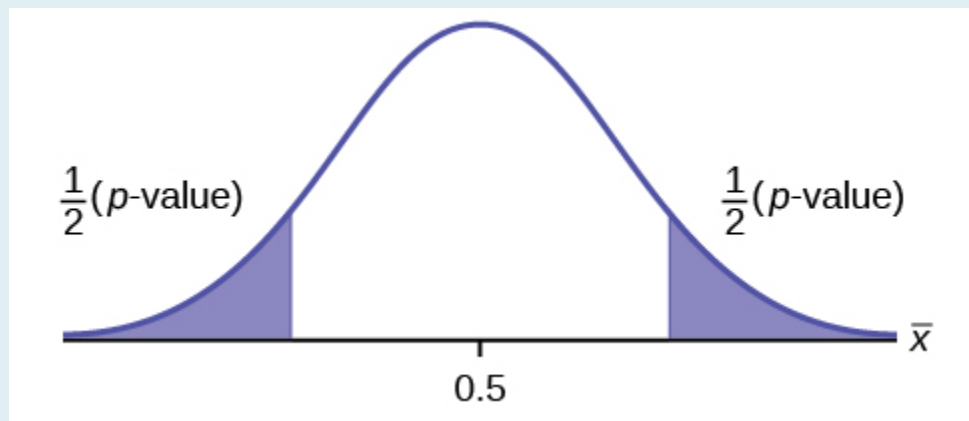
EXAMPLE

Suppose the hypotheses for a hypothesis test are:

$$H_0 : \mu = 0.5$$

$$H_a : \mu \neq 0.5$$

Because the alternative hypothesis is a \neq , this is a two-tailed test. The p – value is the sum of the areas in the two tails of the distribution. Each tail contains exactly half of the p – value.



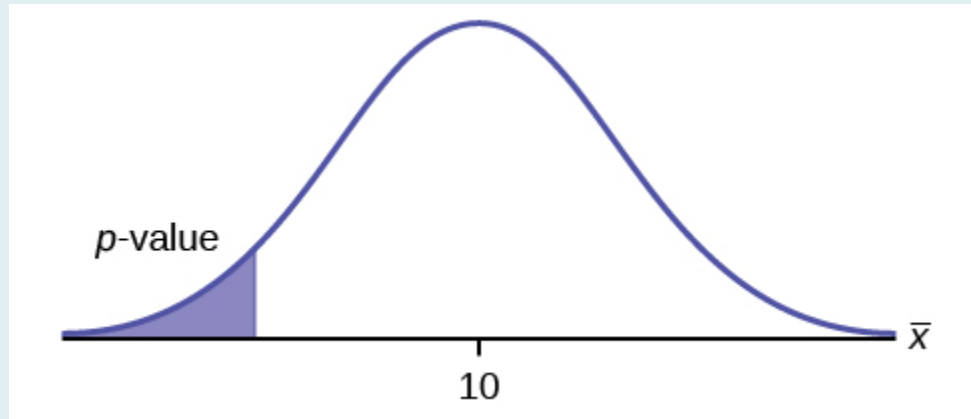
EXAMPLE

Suppose the hypotheses for a hypothesis test are:

$$H_0 : \mu = 10$$

$$H_a : \mu < 10$$

Because the alternative hypothesis is a $<$, this is a left-tailed test. The p – value is the area in the left-tail of the distribution.



Conducting a Hypothesis Test for a Population Mean with a Known Population Standard Deviation

Follow these steps to perform a hypothesis test on a population mean with a known population standard deviation:

1. Write down the null and alternative hypotheses in terms of the population mean μ . Include appropriate units with the values of the mean.
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information for the test and identify the significance level α .
4. When the population standard deviation is **known**, we use a normal distribution with $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ to find the p - value. The p - value is the area in the corresponding tail of the normal distribution.

5. Compare the p - value to the significance level and state the outcome of the test.

- If p - value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
- If p - value $> \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.

6. Write down a concluding sentence specific to the context of the question.

USING EXCEL TO CALCULATE THE p – value FOR A HYPOTHESIS TEST ON A POPULATION MEAN WITH KNOWN POPULATION STANDARD DEVIATION

The p – value for a hypothesis test on a population mean is the area in the tail(s) of the distribution of the sample mean. When the population standard deviation is known, use the normal distribution to find the p – value.

The p – value is the area in the tail(s) of a normal distribution, so the **norm.dist(x,μ,σ,logic operator)** function can be used to calculate the p – value.

- For **x**, enter the value for \bar{x} .
- For **μ**, enter the mean of the sample means μ . Note: Because the test is run assuming the null hypothesis is true, the value for μ is the claim from the null hypothesis.
- For **σ**, enter the standard error of the mean $\frac{\sigma}{\sqrt{n}}$.
- For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

Use the appropriate technique with the **norm.dist** function to find the area in the left-tail or the area in the right-tail.

EXAMPLE

Jeffrey, as an eight-year-old, established a mean time of **16.43** seconds with a standard deviation of **0.8** seconds for swimming the 25-meter freestyle. His dad, Frank, thought that Jeffrey could swim

the 25-meter freestyle faster using goggles. Frank bought Jeffrey a new pair of goggles and timed Jeffrey swimming the 25-meter freestyle 15 different times. In the sample of 15 swims, Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey swim faster than 16.43 seconds. At the 5% significance level, did Jeffrey swim faster wearing the goggles? Assume that the swim times for the 25-meter freestyle are normally distributed.

Solution

Hypotheses:

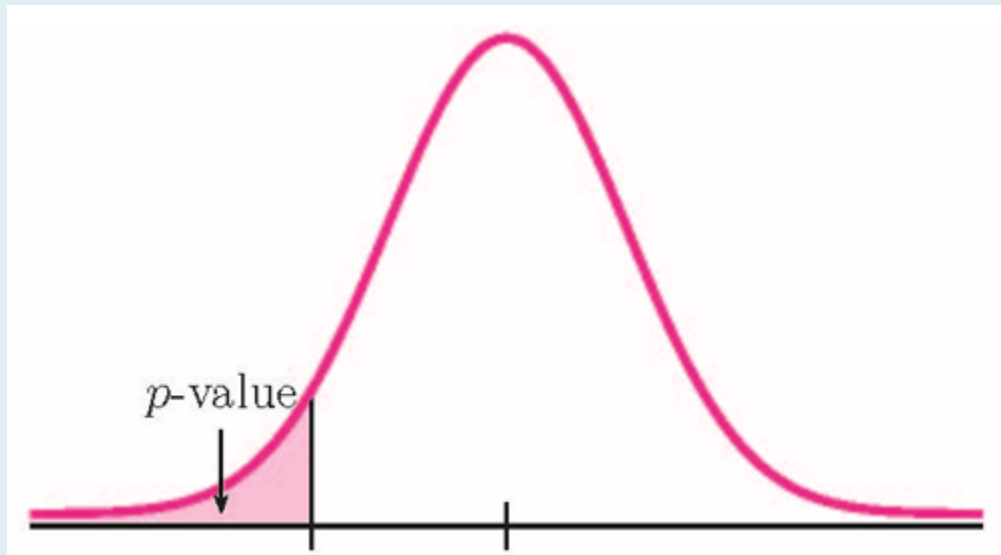
$$H_0 : \mu = 16.43 \text{ seconds}$$

$$H_a : \mu < 16.43 \text{ seconds}$$

p – value :

From the question, we have $n = 15$, $\bar{x} = 16$, $\sigma = 0.8$ and $\alpha = 0.05$.

This is a test on a population mean where the population standard deviation is known ($\sigma = 0.8$). So, we use a normal distribution to calculate the p – value. Because the alternative hypothesis is a $<$, the p – value is the area in the left-tail of the distribution.



Function	norm.dist
Field 1	16
Field 2	16.43
Field 3	0.8/sqrt(15)
Field 4	true
Answer	0.0187

So the p – value = 0.0187.

Conclusion:

Because p – value = 0.0187 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level, there is enough evidence to suggest that Jeffrey's mean swim time with the goggles is less than 16.43 seconds.

NOTES

1. The null hypothesis $\mu = 16.43$ is the claim that Jeffrey's mean swim time with the goggles is 16.43 seconds (the same as it is without the goggles).
2. The alternative hypothesis $\mu < 16.43$ is the claim that Jeffrey's swim time with the goggles is less than 16.43 seconds.
3. The p – value is the area in the left tail of the sampling distribution, to the left of $\bar{x} = 16$. In the calculation of the p – value :
 - The function is **norm.dist** because we are finding the area in the left tail of a normal distribution.
 - Field 1 is the value of \bar{x}
 - Field 2 is the value of μ from the null hypothesis. Remember, we run the test assuming the null hypothesis is true, so that means we assume $\mu = 16.43$.
 - Field 3 is the standard deviation for the sample means $\frac{\sigma}{\sqrt{n}}$. Note that we are **not** using the standard deviation from the population ($\sigma = 0.8$). This is because the p – value is the area under the curve of the distribution of the sample means, not the distribution of the population.

4. The p — value of **0.0187** tells us that under the assumption that Jeffrey's mean swim time with goggles is **16.43** seconds (the null hypothesis), there is only a 1.87% chance that the mean time for the **15** sample swims is **16** seconds or less. This is a small probability, and so it is unlikely to happen, assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis.
5. The Type I error for this problem is to conclude that Jeffrey swims the 25-meter freestyle, on average, in less than **16.43** seconds (the alternative hypothesis) when, in fact, he actually swims the 25-meter freestyle, on average, in **16.43** seconds (the null hypothesis). That is, reject the null hypothesis when the null hypothesis is actually true.
6. The Type II error for this problem is to conclude that Jeffrey swims the 25-meter freestyle, on average, in **16.43** seconds (the null hypothesis) when, in fact, he actually swims the 25-meter freestyle, on average, in less than **16.43** seconds (the alternative hypothesis). That is, do not reject the null hypothesis when the null hypothesis is actually false.

TRY IT

The mean throwing distance of a football for Marco, a high school freshman quarterback, is **40** yards with a standard deviation of **2** yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for **20** throws with the new grip. For the **20** throws, Marco's mean distance was **41.5** yards. The coach thought the different grip helped Marco throw farther than **40** yards. At the 5% significance level, is Marco's mean throwing distance higher with the new grip? Assume the throw distances for footballs are normally distributed.

Click to see Solution

Hypotheses:

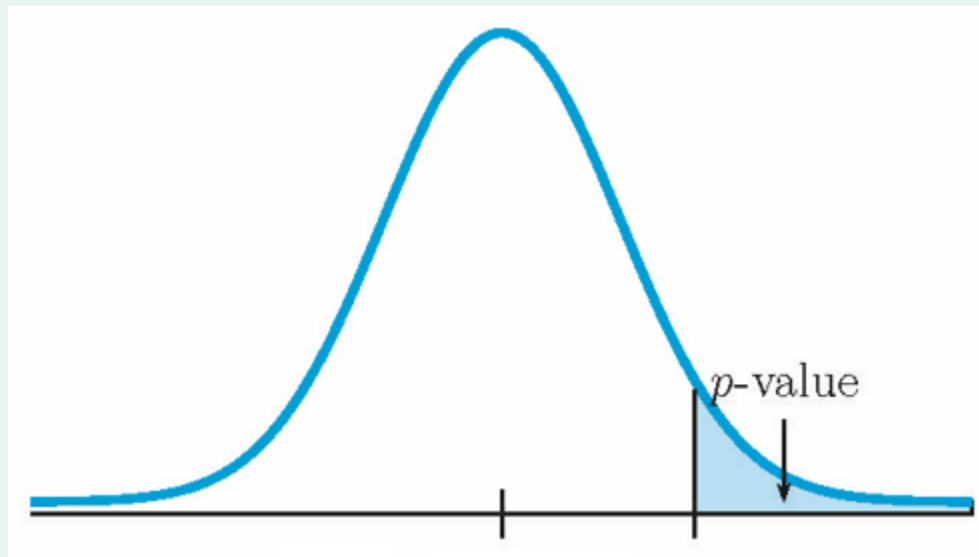
$$H_0 : \mu = 40 \text{ yards}$$

$$H_a : \mu > 40 \text{ yards}$$

p – value:

From the question, we have $n = 20$, $\bar{x} = 41.5$, $\sigma = 2$ and $\alpha = 0.05$.

This is a test on a population mean where the population standard deviation is known ($\sigma = 2$). So we use a normal distribution to calculate the p – value. Because the alternative hypothesis is a $>$, the p – value is the area in the right-tail of the distribution.



Function	1-norm.dist
Field 1	41.5
Field 2	40
Field 3	2/sqrt(20)
Field 4	true
Answer	0.0004

So the p – value = 0.0004.

Conclusion:

Because p – value = 0.0004 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level, there is enough evidence to suggest that Marco's mean throwing distance is greater than 40 yards with the new grip.

NOTES

1. The null hypothesis $\mu = 40$ is the claim that Marco's mean throwing distance with the new grip is 40 yards (the same as it is without the new grip).
2. The alternative hypothesis $\mu > 40$ is the claim that Marco's mean throwing distance with the new grip is greater than 40 yards.
3. The **p — value** is the area in the right tail of the normal distribution. To calculate the area in the right-tail of a normal distribution, we use **1-norm.dist.**
 - Field 1 is the value of \bar{x}
 - Field 2 is the value of μ from the null hypothesis.
 - Field 3 is the standard deviation for the sample means $\frac{\sigma}{\sqrt{n}}$.
4. The **p — value** of 0.0004 tells us that under the assumption that Marco's mean throwing distance with the new grip is 40 yards, there is only a 0.04% chance that the mean throwing distance for the 20 sample throws is more than 40 yards. This is a small probability and so is unlikely to happen, assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis.

EXAMPLE

A local college states in its marketing materials that the average age of its first-year students is 18.3 years with a standard deviation of 3.4 years. But this information is based on old data and does not take into account that more older adults are returning to college. A researcher at the college believes that the average age of its first-year students has changed. The researcher takes a sample of 50 first-

year students and finds the average age is 19.5 years. At the 1% significance level, has the average age of the college's first-year students changed?

Solution

Hypotheses:

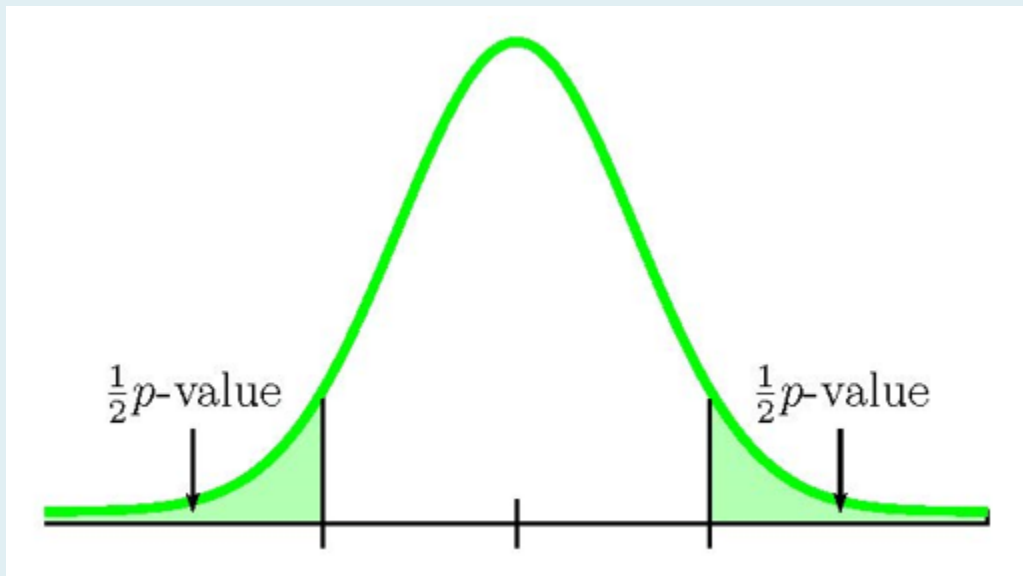
$$H_0 : \mu = 18.3 \text{ years}$$

$$H_a : \mu \neq 18.3 \text{ years}$$

p – value :

From the question, we have $n = 50$, $\bar{x} = 19.5$, $\sigma = 3.4$ and $\alpha = 0.01$.

This is a test on a population mean where the population standard deviation is known ($\sigma = 3.4$). In this case, the sample size is greater than 30. So we use a normal distribution to calculate the p – value. Because the alternative hypothesis is a \neq , the p – value is the sum of the area in the tails of the distribution.



Because there is only one sample, we only have information relating to one of the two tails, either the left tail or the right tail. We need to know if the sample relates to the left tail or right tail because that will determine how we calculate out the area of that tail using the normal distribution. In this case, the sample mean $\bar{x} = 19.5$ is greater than the value of the population mean in the null hypothesis $\mu = 18.3$ ($\bar{x} = 19.5 > 18.3 = \mu$), so the sample information relates to the right-tail of the normal distribution. This means that we will calculate out the area in the right tail using **1-norm.dist**. However, this is a two-tailed test where the p – value is the sum of the area in the

two tails and the area in the right-tail is only one half of the p – value. The area in the left tail equals the area in the right tail and the p – value is the sum of these two areas.

Function	1-norm.dist
Field 1	19.5
Field 2	18.3
Field 3	3.4/sqrt(50)
Field 4	true
Answer	0.0063

So the area in the right tail is 0.0063 and $\frac{1}{2}p$ – value = 0.0063. This is also the area in the left tail, so

$$p$$
 – value = 0.0063 + 0.0063 = 0.0126

Conclusion:

Because p – value = 0.0126 > 0.01 = α , we do not reject the null hypothesis. At the 1\% significance level, there is not enough evidence to suggest that the average age of the college’s first-year students has changed.

NOTES

1. The null hypothesis $\mu = 18.3$ is the claim that the average age of the first-year students is still 18.3 years.
2. The alternative hypothesis $\mu \neq 18.3$ is the claim that the average age of the first-year students has changed from 18.3 years.
3. In a two-tailed hypothesis test that uses the normal distribution, we will only have sample information relating to **one** of the two tails. We must determine which of the tails the sample information belongs to, and then calculate out the area in that tail. The area in each tail represents exactly half of the p – value, so the p – value is the sum of the areas in the two tails.
 - If the sample mean \bar{x} is less than the population mean μ in the null hypothesis ($\bar{x} < \mu$), then the sample information belongs to the **left tail**.

- We use **norm.dist($\bar{x}, \mu, \sigma / \sqrt{n}, \text{true}$)** to find the area in the left tail. The area in the right tail equals the area in the left tail, so we can find the **p – value** by adding the output from this function to itself.
 - If the sample mean \bar{x} is greater than the population mean μ in the null hypothesis ($\bar{x} > \mu$), then the sample information belongs to the **right tail**.
 - We use **1-norm.dist($\bar{x}, \mu, \sigma / \sqrt{n}, \text{true}$)** to find the area in the right tail. The area in the left tail equals the area in the right tail, so we can find the **p – value** by adding the output from this function to itself.
4. The **p – value** of **0.0126** is a large probability compared to the 1\% significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the claim that the average age of first-year students is **18.3** years is most likely correct.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=204#oembed-1>

Video: “Excel Statistical Analysis 43: Hypothesis Testing: P-value & Critical Value Methods: 1 Tail Upper” by excelisfun [32:48] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*





One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=204#oembed-2>

Video: “Excel Statistical Analysis 44: Hypothesis Testing with Z Distribution, 1 Tail Lower (Left) Test” by excelisfun [10:58] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=204#oembed-3>

Video: “Excel Statistical Analysis 45: Hypothesis Testing with Z Distribution, Two Tail Test Example” by excelisfun [9:56] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A particular brand of tire claims that its deluxe tire averages 50,000 kilometres before it needs to be replaced. A group of owners believe this number is too high. From past studies of this tire, the standard deviation is known to be 8,000 kilometres. A survey of owners of that tire design is conducted. From the 35 tires surveyed, the mean lifespan was 46,500 kilometres. At the 5% significance level, test if the tire average before the tire needs replacing is less than claimed.

Click to see Answer

- Hypotheses: $H_0 : \mu = 50,000 \text{ km}$
 $H_a : \mu < 50,000 \text{ km}$
- p – value = 0.0048
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the

tire average before the tire needs replacing is less than \$50,000.

2. From generation to generation, the mean age when smokers first start to smoke is 19 years. However, the standard deviation of that age remains constant of around 2.1 years. A health researcher wants to know if the mean age when smokers first started smoking has changed. In a survey of 40 smokers of this generation, the sample mean was 18.1 years. At the 1% significance level, test the health researcher's claim.

Click to see Answer

- Hypotheses: $H_0 : \mu = 19 \text{ years}$
 $H_a : \mu \neq 19 \text{ years}$
- $p - \text{value} = 0.0067$
- Conclusion: At the 1% significance level, there is enough evidence to conclude that the mean age when smokers first started smoking has changed.

3. The cost of a daily newspaper varies from city to city. In the past, the mean cost of a daily newspaper was \$1.00 with a standard deviation of \$0.20. But a local newspaper publisher believes that the mean cost of a daily newspaper has increased. In a sample of 30 daily newspapers, the mean cost was \$1.06. At the 1% significance level, has the mean cost of a daily newspaper increased?

Click to see Answer

- Hypotheses: $H_0 : \mu = \$1.00$
 $H_a : \mu > \$1.06$
- $p - \text{value} = 0.0502$
- Conclusion: At the 1% significance level, there is not enough evidence to conclude that the mean cost of a daily newspaper has increased.

4. In the past, the mean salary for managers at fast-food restaurants was \$68,500 with a standard deviation of \$9,200. The manager at a local fast-food restaurant wants to test this claim because she believes the mean salary is higher now. In a sample of 50 fast-food restaurant managers, the mean salary was \$70,600. At the 5% significance level, test the manager's claim.

Click to see Answer

- Hypotheses: $H_0 : \mu = \$68,500$
 $H_a : \mu > \$68,500$
- p – value = 0.0533
- Conclusion: At the 5\% significance level, there is not enough evidence to conclude that the mean salary for managers at fast-food restaurants has increased.

5. A market research firm is studying consumer spending habits on Boxing Day. Based on previous research, the mean amount a consumer spends on Boxing Day is \$490 with a standard deviation of \$107. Due to changes in the economy, consumer spending habits have also changed. In a sample of 200 consumers, the mean amount spent on Boxing Day was \$510. At the 5\% significance level, test the market research firm's claim that the mean amount a consumer spends on Boxing Day has changed.

Click to see Answer

- Hypotheses: $H_0 : \mu = \$490$
 $H_a : \mu \neq \$490$
- p – value = 0.0082
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that the mean amount a consumer spends on Boxing Day has changed.

6. A company claims that the mean time a customer waits on hold when calling the customer service support line is 5 minutes with a standard deviation of 1.36 minutes. The manager of the customer service call center believes that improvements made in training and protocols at the center have lowered the mean customer wait time. The manager takes a sample of 70 days and records the wait times. The mean wait time in the sample is 4.65. At the 5\% significance level, test if the mean time a customer waits on hold is less than the company's claim.

Click to see Answer

- Hypotheses: $H_0 : \mu = 5 \text{ minutes}$
 $H_a : \mu < 5 \text{ minutes}$
- p – value = 0.00157
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that the mean time a customer waits on hold is less than 5 minutes.

7. Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years. Suppose that it is somehow known that the population standard deviation is 1.5 years. At the 5% significance level, determine if the mean length of jail time has increased. Assume the distribution of the jail times is approximately normal.

Click to see Answer

- Hypotheses: $H_0 : \mu = 2.5$ years
 $H_a : \mu > 2.5$ years
- p – value = 0.0446
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the mean length of jail time has increased.

“8.6 Hypothesis Tests for a Population Mean with Known Population Standard Deviation” and “8.9 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

8.5 HYPOTHESIS TESTS FOR A POPULATION MEAN WITH UNKNOWN POPULATION STANDARD DEVIATION

LEARNING OBJECTIVES

- Conduct and interpret hypothesis tests for a population mean with unknown population standard deviation.

Some notes about conducting a hypothesis test:

- The null hypothesis H_0 is always an “equal to.” The null hypothesis is the original claim about the population parameter.
- The alternative hypothesis H_a is a “less than,” “greater than,” or “not equal to.” The form of the alternative hypothesis depends on the context of the question.
- The form of the alternative hypothesis tells us if the test is left-tail, right-tail, or two-tail. The alternative hypothesis is the key to conducting the test and finding the correct p – value.
 - If the alternative hypothesis is a “less than”, then the test is left-tail. The p – value is the area in the left-tail of the distribution.
 - If the alternative hypothesis is a “greater than”, then the test is right-tail. The p – value is the area in the right-tail of the distribution.
 - If the alternative hypothesis is a “not equal to”, then the test is two-tail. The p – value is the sum of the area in the two-tails of the distribution. Each tail represents exactly half of the p – value.
- **Think about the meaning of the p – value.** A data analyst (and anyone else) should have more confidence that they made the correct decision to reject the null hypothesis with a smaller p – value (for example, 0.001 as opposed to 0.04) even if using a significance level

of 0.05. Similarly, for a large p – value such as 0.4, as opposed to a p – value of 0.056 (a significance level of 0.05 is less than either number), a data analyst should have more confidence that they made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

- The significance level must be identified before collecting the sample data and conducting the test. Generally, the significance level will be included in the question. If no significance level is given, a common standard is to use a significance level of 5%.
- An alternative approach for hypothesis testing is to use what is called the **critical value approach**. In this book, we will only use the p – value approach. Some of the videos below may mention the critical value approach, but this approach will not be used in this book.

Conducting a Hypothesis Test for a Population Mean with Unknown Population Standard Deviation

Follow these steps to perform a hypothesis test for a population mean with unknown population standard deviation:

1. Write down the null and alternative hypotheses in terms of the population mean μ . Include appropriate units with the values of the mean.
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information for the test and identify the significance level α .
4. When the population standard deviation is **unknown**, the p – value is the area in the corresponding tail of the t -distribution with:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$df = n - 1$$

5. Compare the p – value to the significance level and state the outcome of the test.
 - If p – value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If p – value $> \alpha$, do not reject H_0 .

- The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.

6. Write down a concluding sentence specific to the context of the question.

USING EXCEL TO CALCULATE THE p – value FOR A HYPOTHESIS TEST ON A POPULATION MEAN WITH UNKNOWN POPULATION STANDARD DEVIATION

The p – value for a hypothesis test on a population mean is the area in the tail(s) of the distribution of the sample mean. When the population standard deviation is unknown, use the t -distribution to find the p – value.

If the p – value is the area in the left-tail:

- Use the **t.dist** function to find the p – value. In the **t.dist(t-score, degrees of freedom, logic operator)** function:
 - For **t-score**, enter the value of t calculated from $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$.
 - For **degrees of freedom**, enter the degrees of freedom for the t -distribution $n - 1$.
 - For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.
- The output from the **t.dist** function is the area under the t -distribution to the left of the entered t -score.
- Visit the Microsoft page for more information about the **t.dist** function.

If the p – value is the area in the right-tail:

- Use the **t.dist.rt** function to find the p – value. In the **t.dist.rt(t-score, degrees of freedom)** function:
 - For **t-score**, enter the value of t calculated from $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$.

- For **degrees of freedom**, enter the degrees of freedom for the t -distribution $n - 1$.
- The output from the **t.dist.rt** function is the area under the t -distribution to the right of the entered t -score.
- Visit the Microsoft page for more information about the **t.dist.rt** function.

If the p – value is the sum of area in the tails:

- Use the **t.dist.2t** function to find the p – value. In the **t.dist.2t(t-score, degrees of freedom)** function:
 - For **t-score**, enter the **absolute value** of t calculated from $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$. Note: In the **t.dist.2t** function, the value of the t -score must be a **positive** number. If the t -score is negative, enter the absolute value of the t -score into the **t.dist.2t** function.
 - For **degrees of freedom**, enter the degrees of freedom for the t -distribution $n - 1$.
- The output from the **t.dist.2t** function is the sum of areas in the tails under the t -distribution.
- Visit the Microsoft page for more information about the **t.dist.2t** function.

EXAMPLE

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the following scores:

Mean Scores				
65	67	66	68	72
65	70	63	63	71

The instructor performs a hypothesis test using a 1% level of significance. The test scores are assumed to be from a normal distribution.

Solution

Hypotheses:

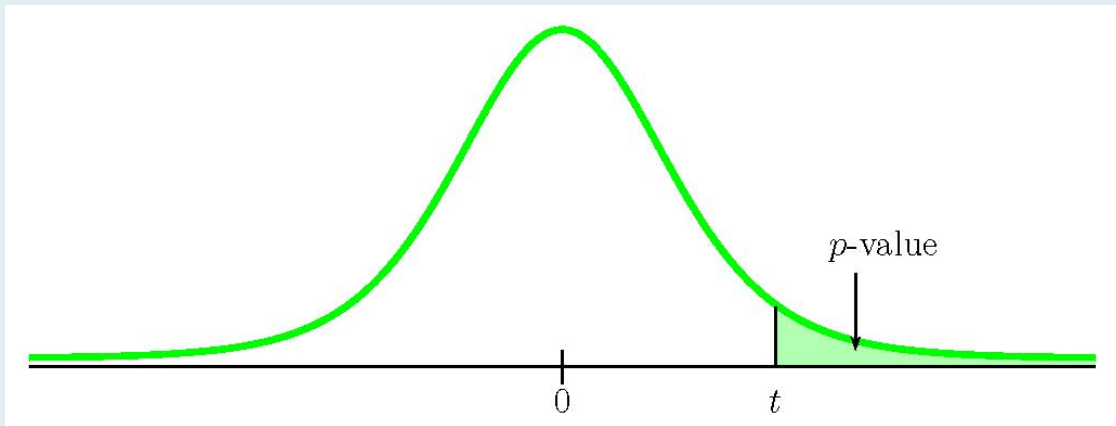
$$H_0 : \mu = 65$$

$$H_a : \mu > 65$$

p – value:

From the question, we have $n = 10$, $\bar{x} = 67$, $s = 3.1972\dots$ and $\alpha = 0.01$.

This is a test on a population mean where the population standard deviation is unknown (we only know the sample standard deviation $s = 3.1972\dots$). So we use a t -distribution to calculate the p – value. Because the alternative hypothesis is a $>$, the p – value is the area in the right-tail of the distribution.



To use the **t.dist.rt** function, we need to calculate out the t -score:

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\ &= \frac{67 - 65}{\frac{3.1972\dots}{\sqrt{10}}} \\ &= 1.9781\dots \end{aligned}$$

The degrees of freedom for the t -distribution is $n - 1 = 10 - 1 = 9$.

Function	t.dist.rt
Field 1	1.9781....
Field 2	9
Answer	0.0396

So the p – value = 0.0396.

Conclusion:

Because p – value = 0.0396 > 0.01 = α , we do not reject the null hypothesis. At the 1\% significance level, there is not enough evidence to suggest that mean score on the test is greater than 65.

NOTES

1. The null hypothesis $\mu = 65$ is the claim that the mean test score is 65.
2. The alternative hypothesis $\mu > 65$ is the claim that the mean test score is greater than 65.
3. Keep all of the decimals throughout the calculation (i.e. in the sample standard deviation, the t -score, etc.) to avoid any round-off error in the calculation of the p – value. This ensures that we get the most accurate value for the p – value.
4. The p – value is the area in the right-tail of the t -distribution, to the right of $t = 1.9781...$
5. The p – value of 0.0396 tells us that under the assumption that the mean test score is 65 (the null hypothesis), there is a 3.96\% chance that the mean test score is 65 or more. Compared to the 1\% significance level, this is a large probability, and so is likely to happen, assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis.

TRY IT

A company claims that the average change in the value of their stock is \$3.50 per week. An investor believes this average is too high. The investor records the changes in the company's stock price over 30 weeks and finds the average change in the stock price is \$2.60 with a standard deviation of \$1.80. At the 5% significance level, is the average change in the company's stock price lower than the company claims?

Click to see Solution

Hypotheses:

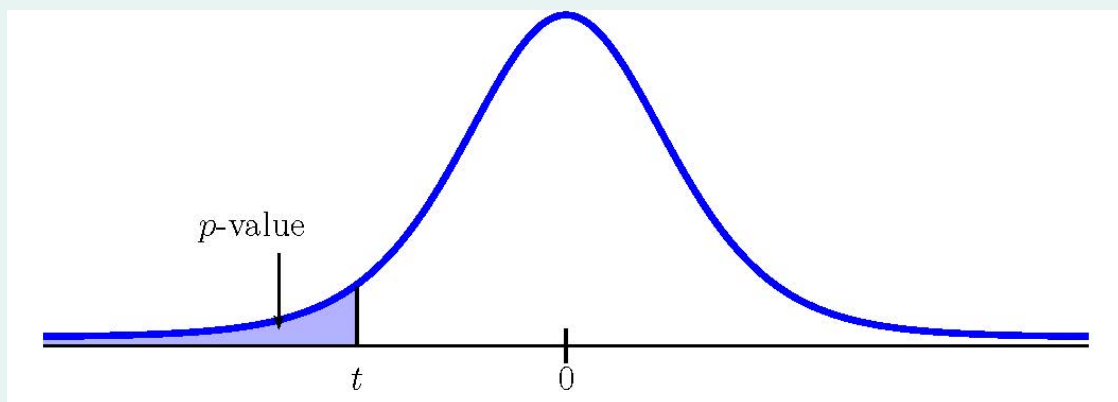
$$H_0 : \mu = \$3.50$$

$$H_a : \mu < \$3.50$$

p – value:

From the question, we have $n = 30$, $\bar{x} = 2.6$, $s = 1.8$ and $\alpha = 0.05$.

This is a test on a population mean where the population standard deviation is unknown (we only know the sample standard deviation $s = 1.8$). So we use a *t*-distribution to calculate the *p* – value. Because the alternative hypothesis is a $<$, the *p* – value is the area in the left-tail of the distribution.



To use the **t.dist** function, we need to calculate out the *t*-score:

$$\begin{aligned}
 t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\
 &= \frac{2.6 - 3.5}{\frac{1.8}{\sqrt{30}}} \\
 &= -1.5699\dots
 \end{aligned}$$

The degrees of freedom for the t -distribution is $n - 1 = 30 - 1 = 29$.

Function	t.dist
Field 1	-1.5699....
Field 2	29
Field 3	true
Answer	0.0636

So the p - value = 0.0636.

Conclusion:

Because p - value = 0.0636 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level, there is not enough evidence to suggest that the average change in the stock price is lower than \$3.50.

NOTES

1. The null hypothesis $\mu = \$3.50$ is the claim that the average change in the company's stock is \$3.50 per week.
2. The alternative hypothesis $\mu < \$3.50$ is the claim that the average change in the company's stock is less than \$3.50 per week.
3. The p - value is the area in the left-tail of the t -distribution, to the left of $t = -1.5699\dots$
4. The p - value of 0.0636 tells us that under the assumption that the average change in the stock is \$3.50 (the null hypothesis), there is a 6.36% chance that the average change is \$3.50 or less. Compared to the 5% significance level, this is a large probability, and so is likely to happen, assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not

reject the null hypothesis. In other words, the company's claim that the average change in their stock price is \$3.50 per week is most likely correct.

EXAMPLE

A paint manufacturer has their production line set-up so that the average volume of paint in a can is 3.78 litres. The quality control manager at the plant believes that something has happened with the production, and the average volume of paint in the cans has changed. The quality control department takes a sample of 100 cans and finds the average volume is 3.62 litres with a standard deviation of 0.7 litres. At the 5% significance level, has the volume of paint in a can changed?

Solution

Hypotheses:

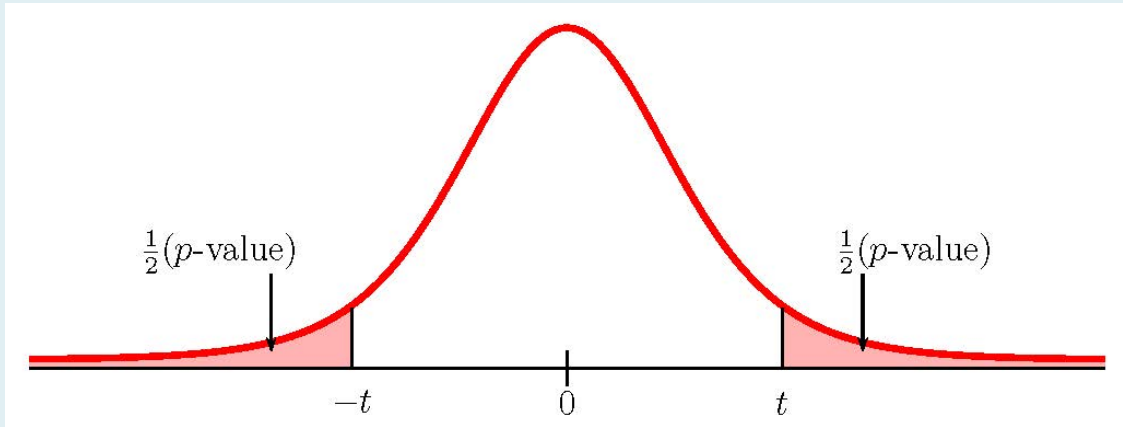
$$H_0 : \mu = 3.78 \text{ liters}$$

$$H_a : \mu \neq 3.78 \text{ liters}$$

p – value:

From the question, we have $n = 100$, $\bar{x} = 3.62$, $s = 0.7$ and $\alpha = 0.05$.

This is a test on a population mean where the population standard deviation is unknown (we only know the sample standard deviation $s = 0.7$). So we use a t -distribution to calculate the p – value. Because the alternative hypothesis is a \neq , the p – value is the sum of the area in the tails of the distribution.



To use the **t.dist.2t** function, we need to calculate out the t -score:

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\ &= \frac{3.62 - 3.78}{\frac{0.07}{\sqrt{100}}} \\ &= -2.2857\dots \end{aligned}$$

The degrees of freedom for the t -distribution is $n - 1 = 100 - 1 = 99$.

Function	t.dist.2t
Field 1	2.2857....
Field 2	99
Answer	0.0244

So the $p - \text{value} = 0.0244$.

Conclusion:

Because $p - \text{value} = 0.0244 < 0.05 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 5\% significance level, there is enough evidence to suggest that the average volume of paint in the cans has changed.

NOTES

1. The null hypothesis $\mu = 3.78$ is the claim that the average volume of paint in the cans is 3.78.
2. The alternative hypothesis $\mu \neq 3.78$ is the claim that the average volume of paint in the cans is not 3.78.
3. Keep all of the decimals throughout the calculation (i.e. in the t -score) to avoid any round-off error in the calculation of the p – value. This ensures that we get the most accurate value for the p – value.
4. The p – value is the sum of the area in the two tails. The output from the **t.dist.2t** function is exactly the sum of the area in the two tails, and so is the p – value required for the test. No additional calculations are required.
5. The **t.dist.2t** function requires that the value entered for the t -score is **positive**. A negative t -score entered into the **t.dist.2t** function generates an error in Excel. In this case, the value of the t -score is negative, so we must enter the absolute value of this t -score into field 1.
6. The p – value of 0.0244 is a small probability compared to the significance level, and so is unlikely to happen, assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the average volume of paint in the cans has most likely changed from 3.78 litres.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=209#oembed-1>

Video: “Excel Statistical Analysis 46: Hypothesis Testing with T Distribution, 1 Tail Upper (Right) Test” by excelisfun [11:02] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=209#oembed-2>

Video: “Excel Statistical Analysis 47: Hypothesis Testing with T Distribution, 1 Tail Lower (Left) Test” by excelisfun [7:48] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=209#oembed-3>

Video: “Excel Statistical Analysis 48: Hypothesis Testing with T Distribution, Two Tail Example” by excelisfun [8:54] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. An article in the *San Jose Mercury News* stated that students in the California State University system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is shorter. You conduct a survey of 49 students and obtain a sample mean of 4.1 with a sample standard deviation of 1.2. Does the data support your claim at the 5% level?

Click to see Answer

- Hypotheses: $H_0 : \mu = 4.5$ years
 $H_a : \mu < 4.5$ years
- p – value = 0.0119
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the mean time for students to finish their undergraduate degree is shorter than 4.1 years.

2. The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows:

Sick Days			
12	3	11	8
4	15	8	6

At the 5% significance level, should the personnel team believe that the mean number is ten?

Click to see Answer

- Hypotheses: $H_0 : \mu = 10$ days
 $H_a : \mu \neq 10$ days
- p – value = 0.2996
- Conclusion: At the 5% significance level, there is not enough evidence to conclude that the mean number of sick days is different from 10 days.

3. In 1955, *Life Magazine* reported that a 25-year-old mother of three worked, on average, an 80 hour week (combining employment and at-home work). Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased. 81 women were surveyed, and the sample mean was 83 with a sample standard deviation of 10. Does it appear that the mean work week has increased for women at the 5% level?

Click to see Answer

- Hypotheses: $H_0 : \mu = 80$ hours
 $H_a : \mu > 80$ hours
- p – value = 0.0042
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the mean work week for women has increased.

4. Based on past studies, the average brown trout's I.Q. is 4. A fish biologist believes that the average brown trout's I.Q. is actually different from this claim. The biologist catches 12 brown

trout and determines the I.Q.s as follows:

Trout I.Q.					
8	7	3	5	3	8
5	4	6	4	6	5

At the 5\% significance level, determine if the average brown trout's I.Q. is different than claimed.

Click to see Answer

- Hypotheses: $H_0 : \mu = 4$
 $H_a : \mu \neq 4$
- p – value = 0.0214
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that the average brown trout's I.Q. is different than 4.

5. The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks ten engineering friends in start-ups for the lengths of their mean work weeks. At the 5\% significance level, test if the mean work week for engineers in a start-up is less than 60 hours.

Work Weeks				
70	60	65	60	50
45	55	55	55	55

Click to see Answer

- Hypotheses: $H_0 : \mu = 60$ hours
 $H_a : \mu < 60$ hours
- p – value = 0.1086
- Conclusion: At the 5\% significance level, there is not enough evidence to conclude that the mean work week for engineers in a start-up is less than 60 hours.

6. The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys

56 online students and finds that the sample mean is 27.4 with a standard deviation of 2.1. At the 1% significance level, determine if the mean age for online students is more than 26.6 years.

Click to see Answer

- Hypotheses: $H_0 : \mu = 26.6$ years
 $H_a : \mu > 26.6$ years
- p – value = 0.0031
- Conclusion: At the 1% significance level, there is enough evidence to conclude that the mean age for online students is more than 26.6 years.

7. According to a national study, registered nurses earned an average annual salary of \$69,110. The head of the nurses union at a local hospital wants to know if the average salary for nurses at the hospital is different than the national average. In a sample of 41 registered nurses at the hospital, the sample average was \$71,121 with a sample standard deviation of \$7,489. At the 5% significance level, determine if the average salary for nurses at the hospital is different from the national average.

Click to see Answer

- Hypotheses: $H_0 : \mu = \$69,110$
 $H_a : \mu \neq \$69,110$
- p – value = 0.0933
- Conclusion: At the 5% significance level, there is not enough evidence to conclude that the average salary for nurses at the hospital is different from the national average.

8. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. In a sample of 35 randomly chosen teenagers, the mean was time spent on the phone was 4.95 hours with a standard deviation of 2.0. At the 5% significance level, determine if the mean time teenagers spend on the phone per week has increased.

Click to see Answer

- Hypotheses: $H_0 : \mu = 4.5$ hours
 $H_a : \mu > 4.5$ hours
- p – value = 0.096
- Conclusion: At the 5% significance level, there is not enough evidence to conclude that

the mean time teenagers spend on the phone per week has increased.

9. According to national weather data, the mean amount of summer rainfall for a certain area is 28.8 cm. The local meteorologist doubts this claim. The meteorologist selects 40 locations around the region and finds that the mean amount of summer rainfall is 18.55 cm with a standard deviation of 3.25 cm. At the 5% significance level, determine if the mean amount of summer rainfall for the area is different than claimed.

Click to see Answer

- Hypotheses: $H_0 : \mu = 28.8 \text{ cm}$
 $H_a : \mu \neq 28.8 \text{ cm}$
- $p - \text{value} = 0.0197$
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the mean amount of summer rainfall for the area is different than 28.8 cm.

10. A survey in the *N.Y. Times Almanac* finds the mean commute time (one way) is 25.4 minutes for the 15 largest US cities. The Austin, TX Chamber of Commerce feels that Austin's commute time is less and wants to publicize this fact. In a sample of 37 randomly selected Austin commuters, the mean was 22.1 minutes with a standard deviation of 5.3 minutes. At the 1% significance level, determine if the mean commute time in Austin is less than the mean commute time of the largest US cities.

Click to see Answer

- Hypotheses: $H_0 : \mu = 25.4 \text{ minutes}$
 $H_a : \mu < 25.4 \text{ years}$
- $p - \text{value} = 0.0003$
- Conclusion: At the 1% significance level, there is enough evidence to conclude that the mean commute time in Austin is less than 25.4 minutes.

Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

8.6 HYPOTHESIS TESTS FOR A POPULATION PROPORTION

LEARNING OBJECTIVES

- Conduct and interpret hypothesis tests for a population proportion.

Some notes about conducting a hypothesis test:

- The null hypothesis H_0 is always an “equal to.” The null hypothesis is the original claim about the population parameter.
- The alternative hypothesis H_a is a “less than,” “greater than,” or “not equal to.” The form of the alternative hypothesis depends on the context of the question.
- The form of the alternative hypothesis tells us if the test is left-tail, right-tail, or two-tail. The alternative hypothesis is the key to conducting the test and finding the correct p – value.
 - If the alternative hypothesis is a “less than”, then the test is left-tail. The p – value is the area in the left-tail of the distribution.
 - If the alternative hypothesis is a “greater than”, then the test is right-tail. The p – value is the area in the right-tail of the distribution.
 - If the alternative hypothesis is a “not equal to”, then the test is two-tail. The p – value is the sum of the area in the two-tails of the distribution. Each tail represents exactly half of the p – value.
- **Think about the meaning of the p – value.** A data analyst (and anyone else) should have more confidence that they made the correct decision to reject the null hypothesis with a smaller p – value (for example, 0.001 as opposed to 0.04) even if using a significance level of 0.05. Similarly, for a large p – value such as 0.4, as opposed to a p – value of 0.056 (a significance level of 0.05 is less than either number), a data analyst should have more

confidence that they made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

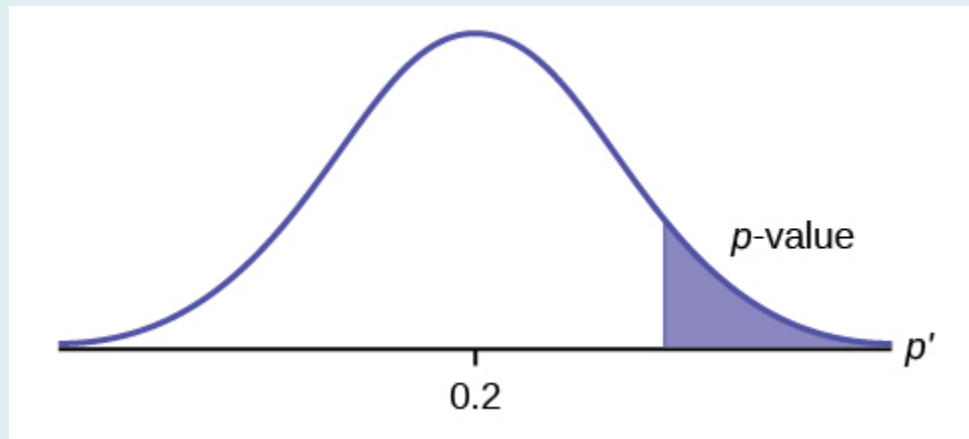
- The significance level must be identified before collecting the sample data and conducting the test. Generally, the significance level will be included in the question. If no significance level is given, a common standard is to use a significance level of 5%.

EXAMPLE

Suppose the hypotheses for a hypothesis test are:

$$\begin{array}{l} H_0: p = 20\% \\ H_a: p > 20\% \end{array}$$

Because the alternative hypothesis is a $>$, this is a right-tail test. The p - value is the area in the right-tail of the distribution.



EXAMPLE

Suppose the hypotheses for a hypothesis test are:

$$\begin{array}{l} H_0: p = 50\% \\ H_a: p \neq 50\% \end{array}$$

Because the alternative hypothesis is a \neq , this is a two-tail test. The p – value is the sum of the areas in the two tails of the distribution. Each tail contains exactly half of the p – value.

EXAMPLE

Suppose the hypotheses for a hypothesis test are:

$$\begin{array}{l} H_0: p = 10\% \\ H_a: p < 10\% \end{array}$$

Because the alternative hypothesis is a $<$, this is a left-tail test. The p – value is the area in the left-tail of the distribution.

Conducting a Hypothesis Test for a Population Proportion

Follow these steps to perform a hypothesis test for a population proportion:

1. Write down the null and alternative hypotheses in terms of the population proportion p . Include appropriate units with the values of the proportion.
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information for the test and identify the significance level.
4. Find the p – value (the area in the corresponding tail) for the test using the appropriate distribution:

- If $n \times p \geq 5$ **and** $n \times (1 - p) \geq 5$, use the normal distribution with $z = \frac{\hat{p} - p}{\sqrt{\frac{p \times (1 - p)}{n}}}$.
- If one of $n \times p < 5$ or $n \times (1 - p) < 5$, use a binomial distribution.

5. Compare the p – value to the significance level and state the outcome of the test.
 - If p – value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If p – value $> \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
6. Write down a concluding sentence specific to the context of the question.

USING EXCEL TO CALCULATE THE p – value FOR A HYPOTHESIS TEST ON A POPULATION PROPORTION

The p – value for a hypothesis test on a population proportion is the area in the tail(s) of distribution of the sample proportion. If both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$, use the normal distribution to find the p – value. If at least one of $n \times p < 5$ or $n \times (1 - p) < 5$, use the binomial distribution to find the p – value.

If both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$:

- The p – value is the area in the tail(s) of a normal distribution, so use the **norm.dist(x, μ , σ , logic operator)** function to calculate the p – value.
 - For **x**, enter the value for \hat{p} .
 - For **μ** , enter the mean of the sample proportions p . Note: Because the test is run assuming the null hypothesis is true, the value for p is the claim from the null hypothesis.
 - For **σ** , enter the standard error of the proportions $\sqrt{\frac{p \times (1 - p)}{n}}$.
 - For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

- Use the appropriate technique with the **norm.dist** function to find the area in the left-tail or the area in the right-tail.

If at least one of $n \times p < 5$ or $n \times (1 - p) < 5$:

- The p – value is found using the binomial distribution.
- If the alternative hypothesis is a $<$, the p – value is the probability of getting at most x successes in n trials where the probability of success is the claim about the population proportion p in the null hypothesis.
 - The p – value is the output from the **binom.dist(x,n,p,logic operator)** function:
 - For **x**, enter the number of successes.
 - For **n**, enter the sample size.
 - For **p**, enter the value of the population proportion p from the null hypothesis.
 - For the **logic operator**, enter **true**. Note: Because we are calculating an at most probability, the logic operator is always true.
- If the alternative hypothesis is a $>$, the p – value is the probability of getting at least x successes in n trials where the probability of success is the claim about the population proportion p in the null hypothesis.
 - The p – value is the output from the **1-binom.dist(x-1,n,p,logic operator)** function:
 - For **x**, enter the number of successes.
 - For **n**, enter the sample size.
 - For **p**, enter the value of the population proportion p in the null hypothesis.
 - For the **logic operator**, enter **true**. Note: Because we are calculating an at least probability, the logic operator is always true.

EXAMPLE

Marketers believe that 92% of adults own a cell phone. A cell phone manufacturer believes that number is actually lower. In a sample of 200 adults, 87% own a cell phone. At the 1% significance level, determine if the proportion of adults that own a cell phone is lower than the marketers' claim.

Solution

Hypotheses:

$$\begin{array}{l} H_0: p = 0.92 \text{ of adults own a cell phone} \\ H_a: p < 0.92 \text{ of adults own a cell phone} \end{array}$$

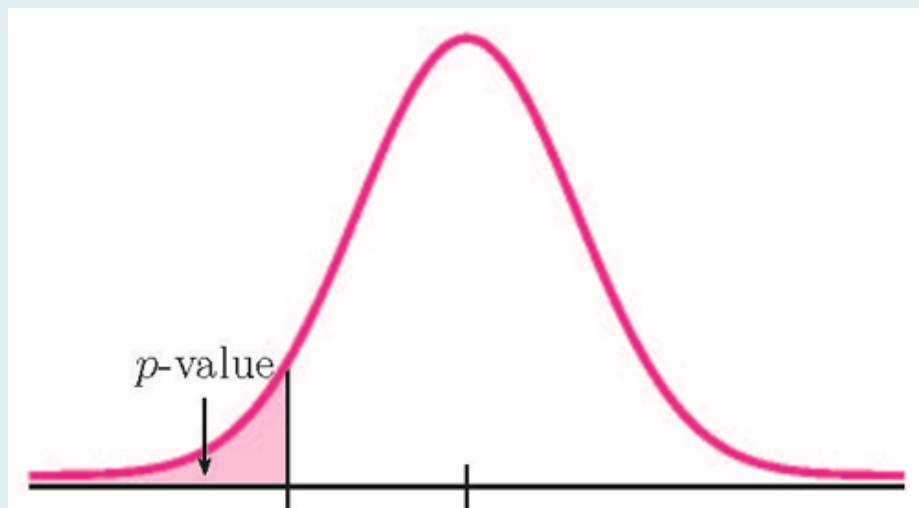
p – value:

From the question, we have $n = 200$, $\hat{p} = 0.87$, and $\alpha = 0.01$.

To determine the distribution, we check $n \times p$ and $n \times (1 - p)$. For the value of p , we use the claim from the null hypothesis ($p = 0.92$).

$$\begin{aligned} n \times p &= 200 \times 0.92 = 184 \geq 5 \\ n \times (1 - p) &= 200 \times (1 - 0.92) = 16 \geq 5 \end{aligned}$$

Because both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$, we use a normal distribution to calculate the p – value. Because the alternative hypothesis is a $<$, the p – value is the area in the left tail of the distribution.



Function	norm.dist
Field 1	0.87
Field 2	0.92
Field 3	$\text{sqrt}(0.92*(1-0.92)/200)$
Field 4	true
Answer	0.0046

So the p – value = 0.0046.

Conclusion:

Because p – value = 0.0046 < 0.01 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 1% significance level there is enough evidence to suggest that the proportion of adults who own a cell phone is lower than 92%.

NOTES

1. The null hypothesis $p=92\%$ is the claim that 92% of adults own a cell phone.
2. The alternative hypothesis $p<92\%$ is the claim that less than 92% of adults own a cell phone.
3. The p – value is the area in the left tail of the sampling distribution, to the left of $\hat{p} = 0.87$. In the calculation of the p – value:
 - The function is **norm.dist** because we are finding the area in the left tail of a normal distribution.
 - Field 1 is the value of \hat{p} .
 - Field 2 is the value of p from the null hypothesis. Remember, we run the test assuming the null hypothesis is true, so that means we assume $p = 0.92$.
 - Field 3 is the standard deviation for the sample proportions $\sqrt{\frac{p \times (1 - p)}{n}}$.
4. The p – value of 0.0046 tells us that under the assumption that 92% of adults own a cell phone (the null hypothesis), there is only a 0.46% chance that the proportion of adults who own a cell phone in a sample of 200 is 87% or less. This is a small probability, and so is unlikely to happen, assuming the null hypothesis is true. This suggests that the assumption

that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the proportion of adults who own a cell phone is most likely less than 92%.

EXAMPLE

A consumer group claims that the proportion of households that have at least three cell phones is 30%. A cell phone company has reason to believe that the proportion of households with at least three cell phones is much higher. Before they start a big advertising campaign based on the proportion of households that have at least three cell phones, they want to test their claim. Their marketing people survey 150 households with the result that 54 of the households have at least three cell phones. At the 1% significance level, determine if the proportion of households that have at least three cell phones is more than 30%.

Solution

Hypotheses:

$$\begin{array}{l} H_0: p = 0.30 \text{ of household have at least 3 cell phones} \\ H_a: p > 0.30 \text{ of household have at least 3 cell phones} \end{array}$$

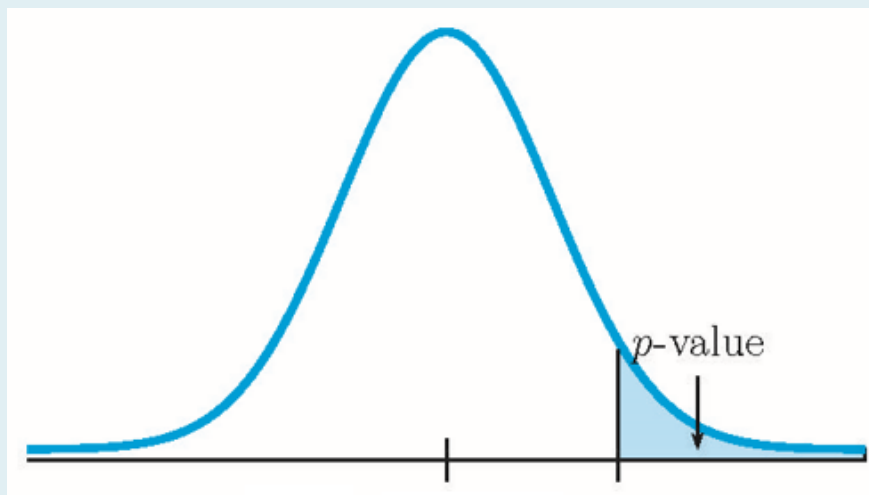
p – value:

From the question, we have $n = 150$, $\hat{p} = \frac{54}{150} = 0.36$, and $\alpha = 0.01$.

To determine the distribution, we check $n \times p$ and $n \times (1 - p)$. For the value of p , we use the claim from the null hypothesis ($p = 0.3$).

$$\begin{aligned} n \times p &= 150 \times 0.3 = 45 \geq 5 \\ n \times (1 - p) &= 150 \times (1 - 0.3) = 105 \geq 5 \end{aligned}$$

Because both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$, we use a normal distribution to calculate the p – value. Because the alternative hypothesis is a $>$, the p – value is the area in the right tail of the distribution.



Function	1-norm.dist
Field 1	0.36
Field 2	0.3
Field 3	sqrt(0.3*(1-0.3)/150)
Field 4	true
Answer	0.0544

So the p – value = 0.0544.

Conclusion:

Because p – value = 0.0544 $>$ 0.01 = α , we do not reject the null hypothesis. At the 1% significance level, there is not enough evidence to suggest that the proportion of households with at least three cell phones is more than 30%.

NOTES

1. The null hypothesis $p=30\%$ is the claim that 30% of households have at least three cell

phones.

2. The alternative hypothesis $p > 30\%$ is the claim that more than 30% of households have at least three cell phones.
3. The p — **value** is the area in the right tail of the sampling distribution, to the right of $\hat{p} = 0.36$. In the calculation of the p — **value**:
 - The function is **1-norm.dist** because we are finding the area in the right tail of a normal distribution.
 - Field 1 is the value of \hat{p} .
 - Field 2 is the value of p from the null hypothesis. Remember, we run the test assuming the null hypothesis is true, so that means we assume $p = 0.3$.
 - Field 3 is the standard deviation for the sample proportions $\sqrt{\frac{p \times (1 - p)}{n}}$.
4. The p — **value** of **0.0544** tells us that under the assumption that 30% of households have at least three cell phones (the null hypothesis), there is a 5.44% chance that the proportion of households with at least three cell phones in a sample of **150** is 36% or more. Compared to the 1% significance level, this is a large probability, and so is likely to happen, assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the claim that 30% of households have at least three cell phones is most likely correct.

TRY IT

A teacher believes that 70% of students in the class will want to go on a field trip to the local zoo. The students in the class believe the proportion is much higher and ask the teacher to verify her claim. The teacher samples 50 students, and 39 reply that they would want to go to the zoo. At the

5\% significance level, determine if the proportion of students who want to go on the field trip is higher than 70\%.

Click to see Solution

Hypotheses:

$$\begin{array}{l} H_0: p = 0.70 \text{ of students want to go on the field trip} \\ H_a: p > 0.70 \text{ of students want to go on the field trip} \end{array}$$

p – value:

From the question, we have $n = 50$, $\hat{p} = \frac{39}{50} = 0.78$, and $\alpha = 0.05$.

$$\begin{aligned} n \times p &= 50 \times 0.7 = 35 \geq 5 \\ n \times (1 - p) &= 50 \times (1 - 0.7) = 15 \geq 5 \end{aligned}$$

Because both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$ we use a normal distribution to calculate the p – value. Because the alternative hypothesis is a $>$, the p – value is the area in the right tail of the distribution.

Function	1-norm.dist
Field 1	0.78
Field 2	0.7
Field 3	$\text{sqrt}(0.7*(1-0.7)/50)$
Field 4	true
Answer	0.1085

So the p – value = 0.1085.

Conclusion:

Because p – value = 0.1085 $>$ 0.05 = α , we do not reject the null hypothesis. At the 5\% significance level, there is not enough evidence to suggest that the proportion of students who want to go on the field trip is higher than 70\%.

NOTES

1. The null hypothesis $p=70\%$ is the claim that 70% of the students want to go on the field trip.
2. The alternative hypothesis $p>70\%$ is the claim that more than 70% of students want to go on the field trip.
3. The p — value of **0.1085** tells us that under the assumption that 70% of students want to go on the field trip (the null hypothesis), there is a 10.85% chance that the proportion of students who want to go on the field trip in a sample of **50** students is 78% or more. Compared to the 5% significance level, this is a large probability, and so is likely to happen, assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the teacher's claim that 70% of students want to go on the field trip is most likely correct.

EXAMPLE

Joan believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is the same or different from 50%. Joan samples 100 first-time brides and 56 reply that they are younger than their grooms. Use a 5% significance level.

Solution

Hypotheses:

$$\begin{array}{l} H_0: p = 50\% \text{ of first-time brides are younger than the groom} \\ H_a: p \neq 50\% \text{ of first-time brides are younger than the groom} \end{array}$$

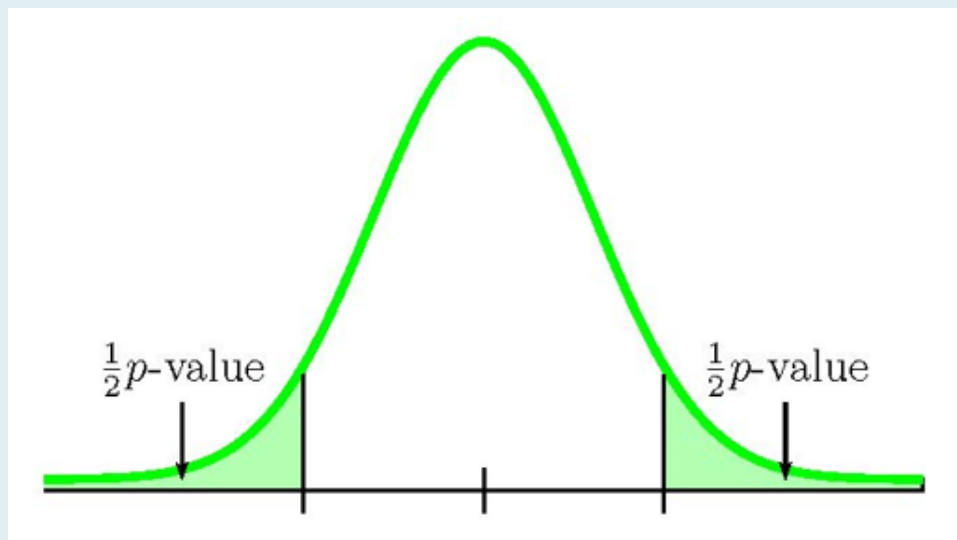
p – value:

From the question, we have $n = 100$, $\hat{p} = \frac{56}{100} = 0.56$, and $\alpha = 0.05$.

To determine the distribution, we check $n \times p$ and $n \times (1 - p)$. For the value of p , we use the claim from the null hypothesis ($p = 0.5$).

$$\begin{aligned} n \times p &= 100 \times 0.5 = 50 \geq 5 \\ n \times (1 - p) &= 100 \times (1 - 0.5) = 50 \geq 5 \end{aligned}$$

Because both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$, we use a normal distribution to calculate the p – value. Because the alternative hypothesis is a \neq , the p – value is the sum of the area in the tails of the distribution.



Because there is only one sample, we only have information relating to one of the two tails, either the left or the right. We need to know if the sample relates to the left or right tail because that will determine how we calculate out the area of that tail using the normal distribution. In this case, the sample proportion $\hat{p} = 0.56$ is greater than the value of the population proportion in the null hypothesis $p = 0.5$ ($\hat{p} = 0.56 > 0.5 = p$), so the sample information relates to the right-tail of the normal distribution. This means that we will calculate out the area in the right tail using **1-norm.dist**. However, this is a two-tailed test where the p – value is the sum of the area in the two tails, and the area in the right-tail is only one half of the p – value. The area in the left tail equals the area in the right tail, and the p – value is the sum of these two areas.

Function	1-norm.dist
Field 1	0.56
Field 2	0.5
Field 3	sqrt(0.5*(1-0.5)/100)
Field 4	true
Answer	0.1151

So the area in the right tail is 0.1151 and $\frac{1}{2}p - \text{value} = 0.1151$. This is also the area in the left tail, so

$$p - \text{value} = 0.1151 + 0.1151 = 0.2302$$

Conclusion:

Because $p - \text{value} = 0.2302 > 0.05 = \alpha$, we do not reject the null hypothesis. At the 5% significance level, there is not enough evidence to suggest that the proportion of first-time brides that are younger than the groom is different from 50%.

NOTES

1. The null hypothesis $p=50\%$ is the claim that the proportion of first-time brides that are younger than the groom is 50%.
2. The alternative hypothesis $p \neq 50\%$ is the claim that the proportion of first-time brides that are younger than the groom is different from 50%.
3. In a two-tailed hypothesis test that uses the normal distribution, we will only have sample information relating to **one** of the two tails. We must determine which of the tails the sample information belongs to, and then calculate out the area in that tail. The area in each tail represents exactly half of the $p - \text{value}$, so the $p - \text{value}$ is the sum of the areas in the two tails.
 - If the sample proportion \hat{p} is less than the population proportion p in the null hypothesis ($\hat{p} < p$), the sample information belongs to the **left tail**.
 - We use **norm.dist($\hat{p}, p, \text{sqrt}(p * (1 - p)/n), \text{true})$** to find the area in the left tail. The area in the right tail equals the area in the left tail, so we can find

the p — value by adding the output from this function to itself.

- If the sample proportion \hat{p} is greater than the population proportion p in the null hypothesis ($\hat{p} > p$), the sample information belongs to the **right tail**.
 - We use `1-norm.dist(\hat{p} , p ,sqrt($p * (1 - p)/n$),true)` to find the area in the right tail. The area in the left tail equals the area in the right tail, so we can find the p — value by adding the output from this function to itself.
4. The p — value of **0.2302** is a large probability compared to the 5% significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the claim that the proportion of first-time brides who are younger than the groom is most likely correct.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=213#oembed-1>

Video: “Excel Statistical Analysis 49: Hypothesis Testing for Proportion (Binominal) using Normal Curve” by excelisfun [7:27] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

EXAMPLE

An online retailer believes that 93% of the visitors to its website will make a purchase. A researcher in the marketing department thinks the actual percentage is lower than claimed. The researcher examines a sample of 50 visits to the website and finds that 45 of the visits resulted in a purchase. At the 1% significance level, determine if the proportion of visits to the website that result in a purchase is lower than claimed.

Solution

Hypotheses:

$$\begin{array}{l} H_0: p = 0.93 \text{ (93\% of visitors make a purchase)} \\ H_a: p < 0.93 \text{ (less than 93\% of visitors make a purchase)} \end{array}$$

p – value:

From the question, we have $n = 50$, $x = 45$, and $\alpha = 0.01$.

To determine the distribution, we check $n \times p$ and $n \times (1 - p)$. For the value of p , we use the claim from the null hypothesis ($p = 0.93$).

$$\begin{aligned} n \times p &= 50 \times 0.93 = 46.5 \geq 5 \\ n \times (1 - p) &= 50 \times (1 - 0.93) = 3.5 < 5 \end{aligned}$$

Because $n \times (1 - p) < 5$, we use a binomial distribution to calculate the p – value. Because the alternative hypothesis is a $<$, the p – value is the probability of getting at most 45 successes in 50 trials.

Function	binom.dist
Field 1	45
Field 2	50
Field 3	0.93
Field 4	true
Answer	0.2710

So the p – value = 0.2710.

Conclusion:

Because $p\text{ -- value} = 0.2710 > 0.01 = \alpha$, we do not reject the null hypothesis. At the 1\% significance level there is not enough evidence to suggest that the proportion of visitors who make a purchase is lower than 93\%.

NOTES

1. The null hypothesis $p=93\%$ is the claim that 93\% of visitors to the website make a purchase.
2. The alternative hypothesis $p<93\%$ is the claim that less than 93\% of visitors to the website make a purchase.
3. The $p\text{ -- value}$ is the binomial probability of getting at most 45 successes (the number in the sample with the characteristic of interest) in 50 trials (the sample size) with a probability of success of 93\% (the value of p in the null hypothesis). In the calculation of the $p\text{ -- value}$:
 - The function is **binom.dist** because we are finding the probability of at most 45 successes.
 - Field 1 is the number of successes x .
 - Field 2 is the sample size n .
 - Field 3 is the probability of success p . This is the claim about the population proportion made in the null hypothesis, so that means we assume $p = 0.93$.
4. The $p\text{ -- value}$ of 0.2710 tells us that under the assumption that 93\% of visitors make a purchase (the null hypothesis), there is a 27.10\% chance that the number of visitors in a sample of 50 who make a purchase is 45 or less. This is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the proportion of visitors to the website who make a purchase, adults is most likely 93\%.

EXAMPLE

A drug company claims that only 4% of people who take their new drug experience any side effects from the drug. A researcher believes that the percentage is higher than the drug company's claim. The researcher takes a sample of 80 people who take the drug and finds that 10% of the people in the sample experience side effects from the drug. At the 5% significance level, determine if the proportion of people who experience side effects from taking the drug is higher than claimed.

Solution

Hypotheses:

$$\begin{array}{l} H_0: p = 0.04 \text{ of people experience side} \\ H_a: p > 0.04 \text{ of people experience side effects} \end{array}$$

p – value:

From the question, we have $n = 80$, $\hat{p} = 0.1$, and $\alpha = 0.05$.

To determine the distribution, we check $n \times p$ and $n \times (1 - p)$. For the value of p , we use the claim from the null hypothesis ($p = 0.04$).

$$n \times p = 80 \times 0.04 = 3.2 < 5$$

Because $n \times p < 5$, we use a binomial distribution to calculate the p – value. Because the alternative hypothesis is a $>$, the p – value is the probability of getting at least 8 successes in 80 trials. (Note: In the sample of size 80, 10% have the characteristic of interest, so this means that $80 \times 0.1 = 8$ people in the sample have the characteristic of interest.)

Function	1-binom.dist
Field 1	7
Field 2	80
Field 3	0.04
Field 4	true
Answer	0.0147

So the p – value = 0.0147.

Conclusion:

Because $p\text{-value} = 0.0147 < 0.05 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level, there is enough evidence to suggest that the proportion of people who experience side effects from taking the drug is higher than 4%.

NOTES

1. The null hypothesis $p=4\%$ is the claim that 4% of the people experience side effects from taking the drug.
2. The alternative hypothesis $p>4\%$ is the claim that more than 4% of the people experience side effects from taking the drug.
3. The $p\text{-value}$ is the binomial probability of getting at least 8 successes (the number in the sample with the characteristic of interest) in 80 trials (the sample size) with a probability of success of 4% (the value of p in the null hypothesis). In the calculation of the $p\text{-value}$:
 - The function is **1-binom.dist** because we are finding the probability of at least 8 successes.
 - Field 1 is $x - 1$ where x is the number of successes. In this case, we are using the complement rule to change the probability of at least 8 successes into 1 minus the probability of at most 7 successes.
 - Field 2 is the sample size n .
 - Field 3 is the probability of success p . This is the claim about the population proportion made in the null hypothesis, so that means we assume $p = 0.04$.
4. The $p\text{-value}$ of 0.0147 tells us that under the assumption that 4% of people experience side effects (the null hypothesis), there is a 1.47% chance that the number of people in a sample of 80 who experience side effects is 8 or more. This is a small probability compared to the significance level, and so is unlikely to happen, assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the proportion of people who experience side effects is most likely greater than 4%.

Exercises

1. Your statistics instructor claims that 60\% of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she cannot quite figure out, most people do not believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. At the 5\% significance level, test if the percentage of students who feel enriched after taking the statistics class is less than the instructor's claim.

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p = 60\% \text{ of students feel more enriched} \\ H_a: p < 60\% \text{ of students feel more enriched} \end{array}$$
- p – value = 0.1308
- Conclusion: At the 5\% significance level, there is not enough evidence to conclude that the percentage of students who feel enriched after taking the statistics class is less than 60\%.

2. Toastmasters International cites a report by Gallop Poll that 40\% of people fear public speaking. A student believes that less than 40\% of students at her school fear public speaking. She randomly surveys 361 schoolmates and finds that 135 report they fear public speaking. At the 1\% significance level, test to determine if the percentage of students at the school who fear public speaking is less than 40\%.

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p = 40\% \text{ of students fear public speaking} \\ H_a: p < 40\% \text{ of students fear public speaking} \end{array}$$
- p – value = 0.1563
- Conclusion: At the 1\% significance level, there is not enough evidence to conclude that the percentage of students who fear public speaking is less than 40\%.

3. According to an article in *Bloomberg Businessweek*, New York City's most recent adult

smoking rate is 14%. Suppose that a survey is conducted to determine this year's rate. In a sample of 70 randomly chosen New York City residents, 16 replied that they smoke. At the 5% significance level, determine if the smoking rate in New York City has changed.

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p = 14\% \text{ of New York City residents} \\ H_a: p \neq 14\% \text{ of New York City residents} \end{array}$$
- p – value = 0.0327
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the smoking rate in New York City has changed.

4. According to the Center for Disease Control website, in 2011, 18% of high school students have smoked a cigarette. A statistics class at a local high school wants to determine if the proportion of students who have smoked a cigarette at their high school is higher than this claim. In a sample of 150 students, 24% said they have smoked a cigarette. At the 5% significance level, test the statistics class's claim.

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p = 18\% \text{ of students have} \\ H_a: p > 18\% \text{ of students have smoked} \end{array}$$
- p – value = 0.0279
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the proportion of students at the high school who have smoked a cigarette is greater than 18%.

5. A recent survey in the *N.Y. Times Almanac* indicated that 48.8% of families own stock. A broker wanted to determine if this survey is valid. He surveyed a random sample of 250 families and found that 142 owned some type of stock. At the 1% significance level, determine if the proportion of families who own stock is different than the result claimed by the survey.

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p = 48.8\% \text{ of families own stock} \\ H_a: p \neq 48.8\% \text{ of families own stock} \end{array}$$
- p – value = 0.0114
- Conclusion: At the 1% significance level, there is not enough evidence to conclude that the proportion of families who own stock is different than 48.8%.

6. According to a national driving association, driver error is listed as the cause of approximately 54% of all fatal automobile accidents. A local insurance agent doubts this claim, suspecting the actual percentage is higher. The insurance agent takes a sample of 60 fatal automobile accidents and finds that 65% were caused by driver error. At the 5% significance level, determine if the percentage of automobile accidents caused by driver error is higher than claimed.

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p = 54\% \text{ of automobile accidents caused by driver error} \\ H_a: p > 54\% \text{ of automobile accidents caused by driver error} \end{array}$$
- p – value = 0.0437
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the percentage of automobile accidents caused by driver error is greater than 54%.

7. According to a library association, 67% of patrons borrow books. The director of a local library believes that the proportion of patrons who borrow books at her library is different from this claim. In a sample of 100 patrons of the local library, 57 borrowed books. At the 1% significance level, is the proportion of patrons who borrow books at the local library different from the library association's claim?

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p = 67\% \text{ of patrons borrow books} \\ H_a: p \neq 67\% \text{ of patrons borrow books} \end{array}$$
- p – value = 0.0334
- Conclusion: At the 1% significance level, there is not enough evidence to conclude that

the proportion of patrons who borrow books at the local library is different than 67%.

8. An all-inclusive resort claims that their guest satisfaction rating is 97%. In a sample of 80 guests at the resort, 74 said they were satisfied with their stay. At the 5% significance level, determine if the satisfaction rate of guests at the resort is less than claimed.

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p = 97\% \text{ satisfaction rating} \\ H_a: p < 97\% \text{ satisfaction rating} \end{array}$$
- p – value = 0.0333
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the satisfaction rating of guests at the resort is less than claimed.

9. A drug company claims that only 9% of people who take their new drug experience any side effects. A researcher wants to test this claim. In a sample of 45 people taking the drug, 9 reported side effects. At the 5% significance level, determine if the proportion of people on the drug who experience side effects is more than claimed.

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p = 9\% \text{ of people on the drug experience side effects} \\ H_a: p > 9\% \text{ of people on the drug experience side effects} \end{array}$$
- p – value = 0.0174
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the proportion of people on the drug who experience side effects is more than 9%.

10. A company that produces a product to help people stop smoking claims that 95% of smokers who use the product stop smoking within six months. In a study to test this claim, 30 smokers use the product for six months. At the end of the six months, 25 of the smokers in the study claim they have stopped smoking. At the 1% significance level, test if the proportion of smokers who use the product stop smoking after six months is less than claimed.

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p = 0.95 \text{ of smokers stop smoking} \\ H_a: p < 0.95 \text{ of smokers stop smoking} \end{array}$$
- p – value = 0.0156
- Conclusion: At the 1% significance level, there is not enough evidence to conclude that the proportion of smokers who use the product stop smoking after six months is less than 95%.

11. A more-than-ten-year-old study reported that 10% of consumers purchased something from an online retailer at least once a week. With the increase in e-commerce and online shopping, a researcher on consumer behaviour believes this percentage is higher today. The researcher takes a sample of 40 consumers, and finds that 22.5% of them purchase something from an online retailer at least once a week. At the 5% significance level, determine the percentage of consumers who purchase something from an online retailer at least once a week is higher than the claim made in the old study.

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p = 0.10 \text{ of consumers shop online at least once a week} \\ H_a: p > 0.10 \text{ of consumers shop online at least once a week} \end{array}$$
- p – value = 0.0155
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the percentage of consumers who purchase something online at least once a week is more than 10%.

PART IX

STATISTICAL INFERENCE FOR TWO POPULATIONS

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then the heart attack rate between the two groups is studied over several years. Other examples of studies between two groups include studies that compare various diet and exercise programs or politicians who compare the proportion of individuals from different income brackets who might vote for them.

Previously, we learned to conduct confidence intervals and hypothesis tests on single means and single proportions. We will extend these ideas in this chapter so that we can compare two means or two proportions to each other. The general procedures are similar to any confidence interval or hypothesis test, following the same basic steps we have already learned but just expanded to include the cases of studying two population parameters.

To compare two means or two proportions, we work with two populations. The groups are classified either as **independent** or **matched pairs**. Independent groups consist of two samples that are independent, which means that the sample values selected from one population are not related in any way to sample values selected from the other population. **Matched pairs** consist of two samples that are dependent, which means there is some relationship between the samples selected from the two populations. In this book, independent groups are used for either two population means or two population proportions and matched pairs are for two population means.

CHAPTER OUTLINE

9.1 Statistical Inference for Two Population Means with Known Population Standard Deviations

9.2 Statistical Inference for Two Population Means with Unknown Population Standard Deviations

9.3 Statistical Inference for Matched Samples

9.4 Statistical Inference for Two Population Proportions

“9.1 Introduction to Statistical Inference with Two Populations” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

9.1 STATISTICAL INFERENCE FOR TWO POPULATION MEANS WITH KNOWN POPULATION STANDARD DEVIATIONS

LEARNING OBJECTIVES

- Construct and interpret a confidence interval for two population means with known population standard deviations.
- Conduct and interpret hypothesis tests for two population means with known population standard deviations.

The comparison of two population means is very common. Often, we want to find out if the two populations under study have the same mean or if there is some difference in the two population means. The approach we take when studying two population means depends on whether the samples are **independent** or **matched**. In the case where the samples are independent, we also have to contend with whether or not we know the population standard deviations.

Two populations are **independent** if the sample taken from population 1 is not related in any way to the sample taken from population 2. In this situation, any relationship between the samples or populations is entirely coincidental.

Throughout this section, we will use subscripts to identify the values for the means, sample sizes, and standard deviations for the two populations:

Symbol for:	Population 1	Population 2
Population Mean	μ_1	μ_2
Population Standard Deviation	σ_1	σ_2
Sample Size	n_1	n_2
Sample Mean	\bar{x}_1	\bar{x}_2
Sample Standard Deviation	s_1	s_2

In order to construct a confidence interval or conduct a hypothesis test on the difference in two population means ($\mu_1 - \mu_2$), we need to use the distribution of the difference in the sample means $\bar{x}_1 - \bar{x}_2$.

- The mean of the distribution of the **difference** in the sample means is $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$.
- The standard deviation of the distribution of the **difference** in the sample means is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

- The distribution of the **difference** in the sample means is normal if **one** of the following is true:
 - Both populations are normally distributed.
 - The sample sizes are large enough ($n_1 \geq 30$ and $n_2 \geq 30$).
- Assuming the distribution of the **difference** of the sample means is normal, the z -score is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

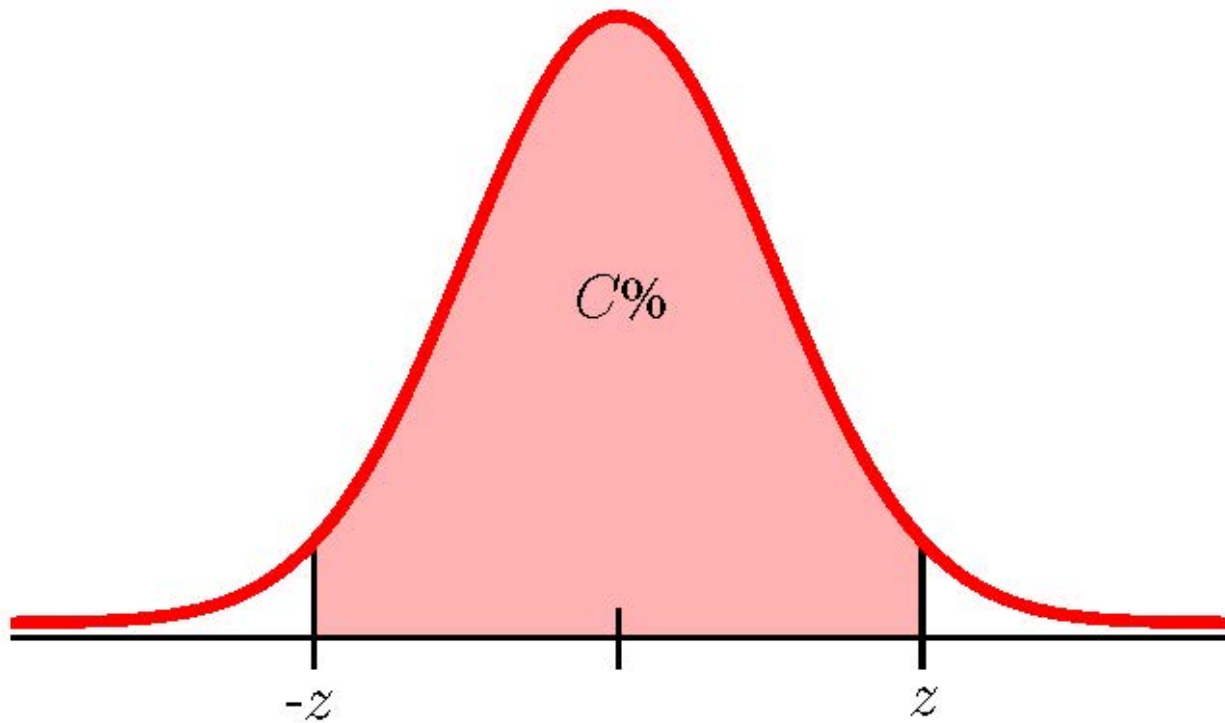
Constructing a Confidence Interval for the Difference in Two Population Means with Known Population Standard Deviation

Suppose a sample of size n_1 with sample mean \bar{x}_1 is taken from population 1 and a sample of size n_2 with sample mean \bar{x}_2 is taken from population 2 where the populations are **independent** and the population standard deviations, σ_1 and σ_2 , are **known**. The limits for the confidence interval with confidence level C for the **difference** in the population means $\mu_1 - \mu_2$ are

$$\text{Lower Limit} = \bar{x}_1 - \bar{x}_2 - z \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\text{Upper Limit} = \bar{x}_1 - \bar{x}_2 + z \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where z is the z -score of the standard normal distribution so that the area to the left of z is $C + \frac{1 - C}{2}$.



NOTE

In order to construct the confidence interval for the difference in two population means with independent samples, we need to check that the distribution of the difference in the sample means follows a normal distribution. This means that we need to check that either the populations are normal or that the sample sizes are large enough (greater than or equal to 30).

CALCULATING THE z -SCORE FOR A CONFIDENCE INTERVAL IN EXCEL

To find the z -score to construct a confidence interval with confidence level C , use the **norm.s.inv(area to the left of z)** function.

- For **area to the left of z**, enter the **entire** area to the left of the z -score you are trying to find.
For a confidence interval, the area to the left of z is $C + \frac{1 - C}{2}$.

The output from the **norm.s.inv** function is the value of the z -score needed to construct the confidence interval.

NOTE

The **norm.s.inv** function requires that we enter the **entire** area to the **left** of the unknown z -score. This area includes the confidence level C (the area in the middle of the distribution) plus the remaining area in the left tail $\frac{1 - C}{2}$.

EXAMPLE

A consumer advocacy group wants to study consumer satisfaction with their shopping experience at the country's two biggest retailers. The group surveyed consumers and asked them to rate one of the retailers in a number of different categories. An overall satisfaction score out of 100 summarized the responses for each consumer sampled. In a sample of 35 consumers for retailer A, the average overall satisfaction score was 79. In a sample of 30 consumers for retailer B, the average overall satisfaction score was 71. Based on prior experience with the satisfaction rating scale, the population standard deviation for retailer A is assumed to be 10, and the population standard deviation for retailer B is assumed to be 12.

1. Construct a 94% confidence interval for the difference in the mean satisfaction score for the two retailers.
2. Interpret the confidence interval found in part 1.
3. Is there evidence to suggest that the mean satisfaction score for retailer A is greater than the mean satisfaction score for retailer B? Explain.

Solution

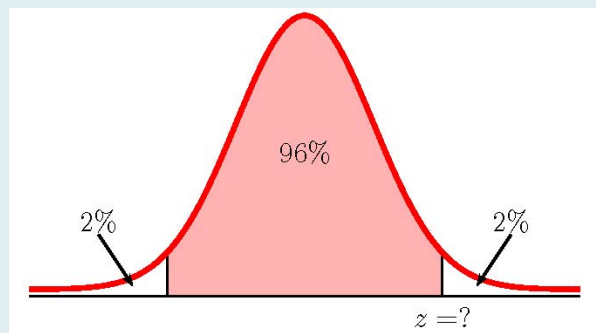
1. Let retailer A be population 1 and retailer B be population 2. These populations are independent because there is no relationship between the consumers sampled for each retailer. From the question, we have the following information:

Retailer A	Retailer B
$n_1 = 35$	$n_2 = 30$
$\bar{x}_1 = 79$	$\bar{x}_2 = 71$
$\sigma_1 = 10$	$\sigma_2 = 12$

The normal distribution applies because both sample sizes are greater than or equal to 30.

To find the confidence interval, we need to find the z -score for the 94% confidence interval.

This means that we need to find the z -score from the standard normal distribution so that the entire area to the left of z is $0.94 + \frac{1 - 0.94}{2} = 0.97$.



Function	norm.s.inv
Field 1	0.97
Answer	1.8807....

So $z = 1.8807 \dots$. The 94% confidence interval is

$$\begin{aligned}
 \text{Lower Limit} &= \bar{x}_1 - \bar{x}_2 - z \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\
 &= 79 - 71 - 1.8807 \dots \times \sqrt{\frac{10^2}{35} + \frac{12^2}{30}} \\
 &= 2.796
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \bar{x}_1 - \bar{x}_2 + z \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\
 &= 79 - 71 + 1.8807 \dots \times \sqrt{\frac{10^2}{35} + \frac{12^2}{30}} \\
 &= 13.204
 \end{aligned}$$

- We are 94% confident that the difference in the mean satisfaction score for the two retailers is between 2.796 and 13.204.
- Because 0 is outside the confidence interval and both limits are positive, it suggests that the difference in the means $\mu_1 - \mu_2$ is greater than 0. That is, $\mu_1 - \mu_2 > 0$ ($\mu_1 > \mu_2$). This suggests that the mean for population 1 (retailer A) is greater than the mean for population 2 (retailer B). So the mean satisfaction score for retailer A is greater than the mean satisfaction

score for retailer B.

NOTES

1. When calculating the limits for the confidence interval, keep all of the decimals in the z -score and other values throughout the calculation. This will ensure that there is no round-off error in the answers. Use Excel to do the calculation of the limits, clicking on the cells containing the z -score or any other values, to ensure that all of the decimal places are used in the calculation.
2. When writing down the interpretation of the confidence interval, make sure to include the confidence level, the actual difference in the population means captured by the confidence interval (i.e. be specific to the context of the question), and appropriate units for the limits.

Conducting a Hypothesis Test for the Difference in Two Independent Population Means with Known Population Standard Deviations

Follow these steps to perform a hypothesis test on the difference in two independent population means with known population standard deviations:

1. Write down the null hypothesis that there is **no** difference in the population means:

$$H_0 : \mu_1 - \mu_2 = 0$$

The null hypothesis is always the claim that the two population means are equal ($\mu_1 = \mu_2$).

2. Write down the alternative hypotheses in terms of the difference in the population means.
The alternative hypothesis will be one of the following:

$$H_a : \mu_1 - \mu_2 < 0 \ (\mu_1 < \mu_2)$$

$$H_a : \mu_1 - \mu_2 > 0 \ (\mu_1 > \mu_2)$$

$$H_a : \mu_1 - \mu_2 \neq 0 \ (\mu_1 \neq \mu_2)$$

3. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
4. Collect the sample information for the test and identify the significance level.
5. Assuming the population standard deviations are known, use the normal distribution to find the p – value. The p – value is the area in the corresponding tail of the normal distribution. The z -score is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

6. Compare the p – value to the significance level and state the outcome of the test.
 - If p – value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If p – value $> \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
7. Write down a concluding sentence specific to the context of the question.

USING EXCEL TO CALCULATE THE p – value FOR A HYPOTHESIS TEST ON TWO INDEPENDENT POPULATION MEANS WITH KNOWN POPULATION STANDARD DEVIATIONS

Assuming that the population standard deviations are known, the p – value for a hypothesis test

on the difference in two independent population means is the area in the tail(s) of the normal distribution.

The p – value is the area in the tail(s) of a normal distribution, so the **norm.dist(x,μ,σ,logic operator)** function can be used to calculate the p – value.

- For x , enter the value for $\bar{x}_1 - \bar{x}_2$.
- For μ , enter 0, the value of $\mu_1 - \mu_2$ from the null hypothesis. This is the mean of the distribution of the differences in the sample means.
- For σ , enter the value of $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, the standard deviation of the distribution of the differences in the sample mean.
- For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

Use the appropriate technique with the **norm.dist** function to find the area in the left-tail, the area in the right-tail or the sum of the area in tails.

EXAMPLE

A floor cleaning company has been using Wax 1 to wax floors for a long time. A new floor wax, Wax 2, has recently come on the market with the claim that it is longer lasting than Wax 1. The company wants to investigate this claim. The company waxed a sample of 20 floors with Wax 1 and found the average number of months the wax lasted was 2.7 months. The company waxed a sample of 20 floors with Wax 2 and found the average number of months the wax lasted was 2.9 months. Based on previous information, the standard deviation for the length of time Wax 1 lasts is 0.33 months, and the standard deviation for the length of time Wax 2 lasts is 0.36 months. Both populations have normal distributions. At the 5% significance level, test if Wax 2 lasts longer, on average, than Wax 1.

Solution

Let Wax 1 be population 1, and Wax 2 be population 2. These populations are independent because

there is no relationship between the length of time each type of wax lasts. From the question, we have the following information:

Wax 1	Wax 2
$n_1 = 20$	$n_2 = 20$
$\bar{x}_1 = 2.7$	$\bar{x}_2 = 2.9$
$\sigma_1 = 0.33$	$\sigma_2 = 0.36$

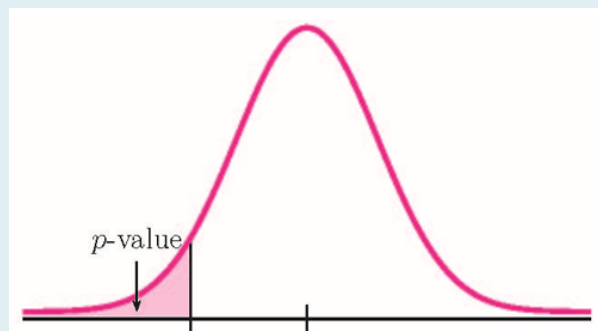
Hypotheses:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 < 0$$

p – value:

This is a test on the difference in two population means where the population standard deviation is known. So, we use a normal distribution to calculate the p – value. Because the alternative hypothesis is a $<$, the p – value is the area in the left-tail of the distribution.



Function	norm.dist
Field 1	2.7-2.9
Field 2	0
Field 3	$\text{sqrt}(0.33^2/20+0.36^2/20)$
Field 4	true
Answer	0.0335

So the p – value = 0.0335.

Conclusion:

Because p – value = 0.0335 < 0.05 = α , we reject the null hypothesis in favour of the

alternative hypothesis. At the 5% significance level, there is enough evidence to suggest that Wax 2 lasts longer than Wax 1.

NOTES

1. The null hypothesis $\mu_1 - \mu_2 = 0$ is the claim that the mean number of months for Wax 1 equals the mean number of months for Wax 2 ($\mu_1 = \mu_2$). That is, the two types of waxes have the same mean.
2. The alternative hypothesis $\mu_1 - \mu_2 < 0$ is the claim that the mean for Wax 1 is less than the mean for Wax 2 ($\mu_1 < \mu_2$). This is the same as saying that the mean for Wax 2 is larger than the mean for Wax 1.
3. The **p-value** is the area in the left tail of the normal distribution. In the calculation of the **p-value**:

- The function is **norm.dist** because we are finding the area in the left tail of a normal distribution.
- Field 1 is the value of $\bar{x}_1 - \bar{x}_2 = 2.7 - 2.9$.
- Field 2 is **0**, the value of $\mu_1 - \mu_2$ from the null hypothesis. Remember, we run the test assuming the null hypothesis is true, so that means we assume $\mu_1 - \mu_2 = 0$.
- Field 3 is the standard deviation for the difference in the sample means

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{0.33^2}{20} + \frac{0.36^2}{20}}.$$

4. The **p-value** of **0.0335** is a small probability compared to the significance level, and so is unlikely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the mean number of months for Wax 1 is less than the mean number of months for Wax 2. For the company, this suggests that they should switch to Wax 2 because it is longer lasting than Wax 1.

EXAMPLE

A consumer advocacy group wants to compare the revolutions per minute (RPM) for two different engines. The group believes that Engine A has a higher average RPM than Engine B. In a sample of 40 Engine A's, the sample mean number of RPMs was 1550. In a sample of 30 Engine B's, the sample mean number of RPMs was 1500. Based on previous information, the standard deviation for the RPMs for Engine A is 75, and the standard deviation for Engine B is 65. At the 1% significance level, is the average RPM for Engine A higher than for Engine B?

Solution

Let Engine A be population 1 and Engine B be population 2. These populations are independent because there is no relationship between the RPMs for the two engines. From the questions, we have the following information:

Engine A	Engine B
$n_1 = 40$	$n_2 = 30$
$\bar{x}_1 = 1550$	$\bar{x}_2 = 1500$
$\sigma_1 = 75$	$\sigma_2 = 65$

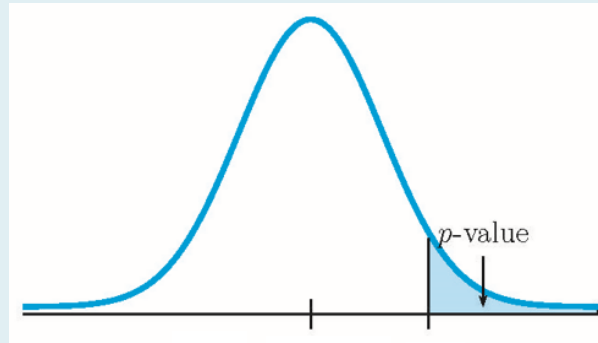
Hypotheses:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 > 0$$

p – value:

This is a test of the difference in two population means where the population standard deviation is known. So, we use a normal distribution to calculate the p – value. Because the alternative hypothesis is a $>$, the p – value is the area in the right tail of the distribution.



Function	1-norm.dist
Field 1	1550-1500
Field 2	0
Field 3	$\text{sqrt}(75^2/40+65^2/30)$
Field 4	true
Answer	0.0014

So the p – value = 0.0014.

Conclusion:

Because p – value = 0.0014 < 0.01 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 1\% significance level, there is enough evidence to suggest that the average RPM for Engine A is higher than for Engine B.

NOTES

1. The null hypothesis $\mu_1 - \mu_2 = 0$ is the claim that the mean RPM for Engine A equals the mean RPM for Engine B ($\mu_1 = \mu_2$). That is, the two engines have the same average RPM.
2. The alternative hypothesis $\mu_1 - \mu_2 > 0$ is the claim that the mean RPM for Engine A is greater than the mean RPM for Engine B ($\mu_1 > \mu_2$).
3. The p – value is the area in the right tail of the normal distribution. In the calculation of the p – value:
 - The function is **1-norm.dist** because we are finding the area in the right tail of a normal distribution.

- Field 1 is the value of $\bar{x}_1 - \bar{x}_2 = 1550 - 1500$.
- Field 2 is 0, the value of $\mu_1 - \mu_2$ from the null hypothesis. Remember, we run the test assuming the null hypothesis is true, so that means we assume $\mu_1 - \mu_2 = 0$.
- Field 3 is the standard deviation for the difference in the sample means

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{75^2}{40} + \frac{65^2}{30}}.$$

4. The p - value of 0.0014 is a small probability compared to the significance level, and so is unlikely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the mean RPM for Engine A is greater than the mean RPM for Engine B, just as the consumer advocacy group claimed.

EXAMPLE

The student union at a local college owns two coffee shops on campus: The Study Cafe and Coffee&Books. The student union wants to find out if there is a difference in the average amount students spend per transaction at each of the coffee shops. In a sample of 65 transactions at the Study Cafe, the average amount spent was \$9.40. In a sample of 50 transactions at Coffee&Books, the average amount spent was \$10.15. Based on previous information, the standard deviation for the amount spent at the Study Cafe is \$1.35 and the standard deviation for Coffee&Books is \$2.70. At the 5% significance level, is there a difference in the average amount spent per transaction at the two coffee shops?

Solution

Let the Study Cafe be population 1, and Coffee&Books be population 2. These populations are

independent because there is no relationship between the amount spent at each coffee shop. From the question, we have the following information:

The Study Cafe	Coffee&Books
$n_1 = 65$	$n_2 = 50$
$\bar{x}_1 = 9.40$	$\bar{x}_2 = 10.15$
$\sigma_1 = 1.35$	$\sigma_2 = 2.70$

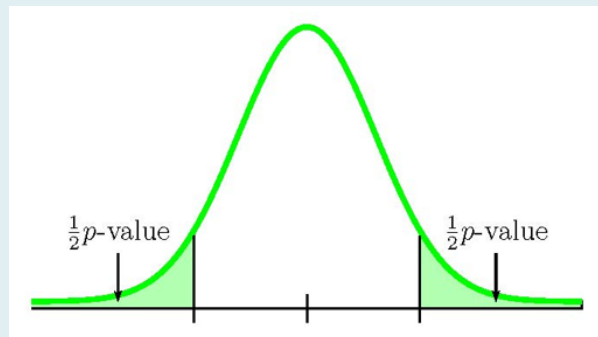
Hypotheses:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

p – value:

This is a test on the difference in two population means where the population standard deviation is known. So, we use a normal distribution to calculate the p – value. Because the alternative hypothesis is a \neq , the p – value is the sum of the area in the two tails of the distribution.



We need to know if the sample information relates to the left or right tail because that will determine how we calculate out the area of that tail using the normal distribution. In this case, the $\bar{x}_1 < \bar{x}_2$ ($9.4 < 10.15$), so the sample information relates to the left tail of the normal distribution. This means that we will calculate out the area in the left tail using **norm.dist**. However, this is a two-tailed test where the p – value is the sum of the area in the two tails, and the area in the left tail is only one-half of the p – value. The area in the left tail equals the area in the right tail, and the p – value is the sum of these two areas.

Function	norm.dist
Field 1	9.40-10.15
Field 2	0
Field 3	$\text{sqrt}(1.35^2/65+2.7^2/50)$
Field 4	true
Answer	0.0360

So the area in the left tail is 0.0360, which means $\frac{1}{2}p - \text{value} = 0.0360$. This is also the area in the right tail, so

$$p - \text{value} = 0.0360 + 0.0360 = 0.0720$$

Conclusion:

Because $p - \text{value} = 0.0720 > 0.05 = \alpha$, we do not reject the null hypothesis. At the 5% significance level, there is not enough evidence to suggest that there is a difference in the average amount spent at the two coffee shops.

NOTES

1. The null hypothesis $\mu_1 - \mu_2 = 0$ is the claim that the mean amount spent at the Study Cafe equals the mean amount spent at Coffee&Books ($\mu_1 = \mu_2$). That is, the average amount spent is the same at both coffee shops.
2. The alternative hypothesis $\mu_1 - \mu_2 \neq 0$ is the claim that the mean amount spent at the Study Cafe is different than the mean amount spent at Coffee&Books ($\mu_1 \neq \mu_2$).
3. In a two-tailed hypothesis test that uses the normal distribution, we will only have sample information relating to **one** of the two tails. We must determine which of the tails the sample information belongs to, and then calculate out the area in that tail. The area in each tail represents exactly half of the $p - \text{value}$, so the $p - \text{value}$ is the sum of the areas in the two tails.
 - If the sample mean \bar{x}_1 is **less than** the sample mean \bar{x}_2 ($\bar{x}_1 < \bar{x}_2$), the sample information belongs to the **left tail**.

- We use **norm.dist** $(\bar{x}_1 - \bar{x}_2, 0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \text{true})$ to find the area in the left tail. The area in the right tail equals the area in the left tail, so we can find the **p — value** by adding the output from this function to itself.
 - If the sample mean \bar{x}_1 is **greater than** the sample mean \bar{x}_2 ($\bar{x}_1 > \bar{x}_2$), the sample information belongs to the **right tail**.
 - We use **1-norm.dist** $(\bar{x}_1 - \bar{x}_2, 0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \text{true})$ to find the area in the right tail. The area in the left tail equals the area in the right tail, so we can find the **p — value** by adding the output from this function to itself.
4. The **p — value** of **0.0720** is a large probability compared to the significance level, and so is likely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the mean amount spent at the Study Cafe equals the mean amount spent at Coffee&Books.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=221#oembed-1>

Video: “Excel 2013 Statistical Analysis #64: Confidence Interval for Population Differences Sigma Known” by excelisfun [9:52] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=221#oembed-2>

Video: “Excel 2013 Statistical Analysis #65: Hypothesis Testing for Population Differences Sigma Known” by excelisfun [16:48] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. You work at company Alpha. You have heard that the mean hourly pay rate of employees at another company, Beta, is higher than at Alpha. In a sample of 42 employees at Alpha, the mean hourly pay rate was \$39.17. In a sample of 38 employees at Beta, the mean hourly pay rate was \$40.27. It is known that the standard deviation of the hourly pay rate at Alpha is \$3.16, and the standard deviation of the hourly pay rate at Beta is \$2.95. At the 5% significance level, determine if the mean hourly pay rate at Alpha is less than the mean hourly pay rate at Beta.

Click to see Answer

Let Alpha be population 1 and Beta be population 2.

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$
 $H_a : \mu_1 - \mu_2 < 0$
- p – value = 0.0537
- Conclusion: At the 5% significance level, there is not enough evidence to conclude that the mean hourly pay rate at Alpha is less than the mean hourly pay rate at Beta.

2. Baseball scouts are evaluating two different pitching prospects: Wesley and Rodriguez. The scouts are interested in the mean speed of each pitcher’s fastball. In a sample of 14 of Wesley’s fastballs, the mean speed was 86 mph. In a sample of 14 of Rodriguez’s fastballs, the mean speed was 91 mph. It is known that the population of all of Wesley’s fastballs is normal with a standard deviation of 3 mph and that the population of all of Rodriguez’s fastballs is normal with a standard deviation of 7. At the 1% significance

level, is there a difference in the mean speed of the two pitchers' fastballs?

Click to see Answer

Let Wesley be population 1, and Rodriquez be population 2.

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$
 $H_a : \mu_1 - \mu_2 \neq 0$
- p - value = 0.0140
- Conclusion: At the 1\% significance level, there is enough evidence to conclude that there is no difference in the mean speed of the two pitchers' fastballs.

3. A researcher is testing the effects of plant food on plant growth. A sample of 9 plants given the plant food, the mean height of the plants after eight weeks was 40 cm. A sample of 9 plants not given the plant food, the mean height of the plants after eight week was 35 cm. It is known that the distribution of the heights of the plants given the plant food is normal with a standard deviation of 6.25 cm and that the distribution of the heights of the plants not given the plant food is normal with a standard deviation of 3.75 cm. At the 5\% significance level, determine if the plants given the plant food have a larger mean height than the plants not given the plant food.

Click to see Answer

Let "given food" be population 1 and "not give food" be population 2.

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$
 $H_a : \mu_1 - \mu_2 > 0$
- p - value = 0.0198
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that the plants given the plant food have a larger mean height than the plants not given the plant food.

4. Two metal alloys, Gamma and Zeta, are being considered as material for ball bearings. The mean melting point of the two alloys is to be compared. In a sample of 45 pieces of Gamma, the mean melting point was $430^\circ C$. In a sample of 60 pieces of Zeta, the mean melting point was $450^\circ C$. It is known that the population standard deviation of the melting point for Gamma is $35^\circ C$ and that the population standard deviation of the melting point for Zeta is $40^\circ C$. At the 1\% significance level, determine if there is a

difference in the mean melting point of the two alloys.

Click to see Answer

Let Gamma be population 1 and Zeta be population 2.

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$
 $H_a : \mu_1 - \mu_2 \neq 0$
- p - value = 0.0064
- Conclusion: At the 1\% significance level, there is enough evidence to conclude that there is a difference in the mean melting point of the two alloys.

5. Parents of teenage boys often complain that auto insurance costs more, on average, for teenage boys than for teenage girls. A group of concerned parents examines a random sample of insurance bills. The mean annual cost for 36 teenage boys was \$679. The mean annual cost for 30 teenage girls was \$593. From past years, it is known that the population standard deviation for each group is \$180. At the 5\% significance level, determine if they believe the mean cost for auto insurance for teenage boys is greater than that for teenage girls.

Click to see Answer

Let boys be population 1 and girls be population 2.

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$
 $H_a : \mu_1 - \mu_2 > 0$
- p - value = 0.0266
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that the mean cost for auto insurance for boys is greater than for girls.

6. The known standard deviation in salary for all mid-level professionals in the financial industry is \$11, 000. Company A and Company B are in the financial industry. In a sample of 30 mid-level professionals in Company A, the sample mean is \$80, 000. In a sample of 35 mid-level professionals in Company B, the sample mean is \$96, 000.
 - a. Construct a 99\% confidence interval for the difference in the mean salary for mid-level professionals at the two companies.
 - b. Interpret the confidence interval in part (a).
 - c. Is it reasonable to claim that mean salary for mid-level professionals the same at the

two companies? Explain.

Click to see Answer

Let Company A be population 1 and Company B be population 2.

- a. Lower Limit = $-23,049.72$, Upper Limit = -8950.28
- b. There is a 99% probability that the difference in the mean salary for mid-level professionals at the two companies is between $-\$23,049.72$ and $-\$8,950.28$.
- c. No. Because both limits are negative, it suggests that the difference in the means $\mu_1 - \mu_2$ is less than 0. That is $\mu_1 - \mu_2 < 0$ or $\mu_1 < \mu_2$. So the mean for Company A is less than the mean for Company B.

7. A group of transfer-bound students wondered if they will spend the same mean amount on texts and supplies each year at their four-year university as they have at their community college. In a sample of 54 students at the community college, the mean amount spent on texts and supplies was \$974. In a sample of 66 students at the four-year university, the mean amount spent on texts and supplies was \$1,011. The population standard deviation for texts and supplies at the community college is known to be \$163, and the population standard deviation for texts and supplies at the four-year university is known to be \$87.
 - a. Construct a 96% confidence interval for the difference in the mean amount students spend on texts and supplies at university and community college.
 - b. Interpret the confidence interval in part (a).
 - c. Is it reasonable to claim that the mean amount students spend on texts and supplies is the same at university and community college? Explain.

Click to see Answer

Let community college be population 1 and university be population 2.

- a. Lower Limit = -87.59 , Upper Limit = 13.59
- b. There is a 96% probability that the difference in the mean amount students spend on texts and supplies at university and community college is between $-\$87.59$ and $\$13.59$.
- c. Yes. Because 0 is inside the confidence interval, it suggests that the difference in the means is 0. That is $\mu_1 - \mu_2 = 0$ or $\mu_1 = \mu_2$. So the mean amount students spend on texts and supplies is the same.

8. A manufacturing company uses two different processes to produce a certain item. The

company wants to study the difference in the mean times the processes take to produce the item. In a sample of 52 items produced using Process 1, the mean completion time was 34 minutes. In a sample of 44 items produced using Process 2, the mean completion time was 31 minutes. It is known that the standard deviation of the completion times for items using Process 1 is 2.7 minutes, and the standard deviation of the completion times for items using Process 2 is 2.1 minutes.

- Construct a 98% confidence interval for the difference in the mean time to produce the item using the two processes.
- Interpret the confidence interval in part (a).
- Is it reasonable to claim that the mean completion time for Process 1 is greater than for Process 2? Explain.

Click to see Answer

Let Process 1 be population 1 and Process 2 be population 2.

- Lower Limit = 1.86, Upper Limit = 4.14
- There is a 98% probability that the difference in the mean the mean time to produce the item using the two processes is between 1.86 minutes and 4.14 minutes.
- Yes. Because both limits are positive, it suggests that the difference in the means is greater than 0. That is $\mu_1 - \mu_2 > 0$ or $\mu_1 > \mu_2$. So, the mean completion time for Process 1 is greater than for Process 2.

“9.2 Statistical Inference for Two Population Means with Known Population Standard Deviations” and “9.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

9.2 STATISTICAL INFERENCE FOR TWO POPULATION MEANS WITH UNKNOWN POPULATION STANDARD DEVIATIONS

LEARNING OBJECTIVES

- Construct and interpret a confidence interval for two population means with unknown population standard deviations.
- Conduct and interpret hypothesis tests for two population means with unknown population standard deviations.

The comparison of two population means is very common. Often, we want to find out if the two populations under study have the same mean or if there is some difference between the two population means. The approach we take when studying two population means depends on whether the samples are **independent** or **matched**. In the case the samples are independent, we also have to contend with whether or not we know the population standard deviations.

Two populations are **independent** if the sample taken from population 1 is not related in any way to the sample taken from population 2. In this situation, any relationship between the samples or populations is entirely coincidental.

Throughout this section, we will use subscripts to identify the values for the means, sample sizes, and standard deviations for the two populations:

Symbol for:	Population 1	Population 2
Population Mean	μ_1	μ_2
Population Standard Deviation	σ_1	σ_2
Sample Size	n_1	n_2
Sample Mean	\bar{x}_1	\bar{x}_2
Sample Standard Deviation	s_1	s_2

In order to construct a confidence interval or conduct a hypothesis test on the difference in two population means ($\mu_1 - \mu_2$), we need to use the distribution of the difference in the sample means $\bar{x}_1 - \bar{x}_2$.

- The mean of the distribution of the **difference** in the sample means is $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$.
- The standard deviation of the distribution of the **difference** in the sample means is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

- The distribution of the **difference** in the sample means is normal if **one** of the following is true:
 - Both populations are normally distributed.
 - The sample sizes are large enough ($n_1 \geq 30$ and $n_2 \geq 30$).
- Assuming the distribution of the **difference** of the sample means is normal, the z -score is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

As we have seen previously when working with confidence intervals and hypothesis testing for a single population mean, when the population standard deviation is unknown, and we must use the sample standard deviation as an estimate for the population standard deviation, we use a t -distribution. We do the same thing when working with the two population means. When the population standard deviations are unknown, we use the sample standard deviations as estimates for the population standard deviations σ_1 and σ_2 . In this situation, we use a t -distribution for the distribution of the difference in the sample means. So, when the population standard deviations are unknown for a confidence interval or hypothesis test on the difference in two population means, we will use a t -distribution. The t -score and the degrees of freedom are:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2}$$

Obviously, the degrees of freedom formula is somewhat complicated. But a computer makes the calculation a bit more manageable. The output from the degrees of freedom formula is rarely a whole number. After calculating the value of df using the above formula, round the output from this formula **down** to the next whole number to get the required degrees of freedom for the t -distribution.

Constructing a Confidence Interval for the Difference in Two Population Means with Unknown Population Standard Deviations

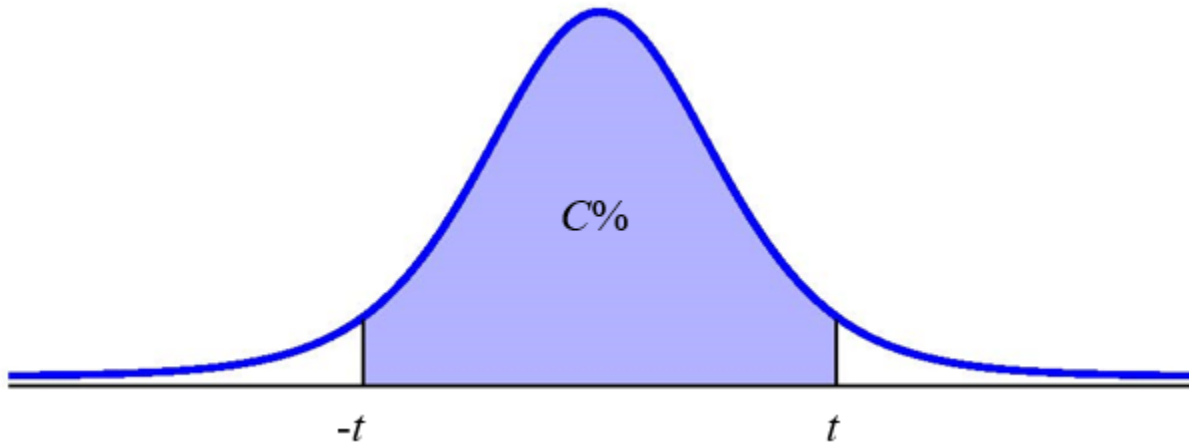
Suppose a sample of size n_1 with sample mean \bar{x}_1 and standard deviation s_1 is taken from population 1 and a sample of size n_2 with sample mean \bar{x}_2 and standard deviation s_2 is taken from population 2 where the populations are independent and the population standard deviations are **unknown**. The limits for the confidence interval with confidence level C for the difference in the population means $\mu_1 - \mu_2$ are:

$$\text{Lower Limit} = \bar{x}_1 - \bar{x}_2 - t \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{Upper Limit} = \bar{x}_1 - \bar{x}_2 + t \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t is the positive t -score of the t -distribution with

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2}$$
 so that the area under the curve in between $-t$ and t is $C\%$.



NOTES

1. In order to construct the confidence interval for the difference in two population means with independent samples, we need to check that the distribution of the difference in the sample means follows a normal distribution. This means that we need to check that either the populations are normal or that the sample sizes are large enough (greater than or equal to 30).
2. When the population standard deviations are unknown, we must use a t -distribution in the construction of the confidence interval.
3. The value of degrees of freedom must be a whole number. After using the formula provided above, remember to round the value **down** to the next whole number to get the required degrees of freedom for the t -distribution.

CALCULATING THE t -SCORE FOR A CONFIDENCE INTERVAL IN

EXCEL

To find the t -score to construct a confidence interval with confidence level C , use the **t.inv.2t(area in the tails, degrees of freedom)** function.

- For **area in the tails**, enter the **sum** of the area in the tails of the t -distribution. For a confidence interval, the area in the tails is $1 - C$.
- For **degrees of freedom**, enter the degrees of freedom calculated using

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2} \right)^2}.$$

The output from the **t.inv.2t** function is the value of t -score needed to construct the confidence interval.

NOTE

1. The **t.inv.2t** function requires that we enter the **sum** of the area in **both** tails. The area in the middle of the distribution is the confidence level C , so the sum of the area in both tails is the leftover area $1 - C$.
2. The degrees of freedom for a t -distribution **must** be a **whole number**. The output from the

degrees of freedom formula $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2} \right)^2}$ is almost

never a whole number. After calculating the value of df using the formula, **round the value down to the next whole number**. Remember to enter the rounded down value of df for the degrees of freedom in the **t.inv.2t** function.

EXAMPLE

A company that manufactures and services photocopiers wants to study the difference in the average repair time for the two different models of photocopiers they make. In a sample of 60 repairs of photocopier A, the mean repair time was 84.2 minutes with a standard deviation of 19.4 minutes. In a sample of 70 repairs of photocopier B, the mean repair time was 91.6 minutes with a standard deviation of 18.8 minutes.

1. Construct a 95% confidence interval for the difference in the mean repair time for the two photocopiers.
2. Interpret the confidence interval found in part 1.
3. Is there evidence to suggest that the mean repair times for the photocopiers is the same? Explain.

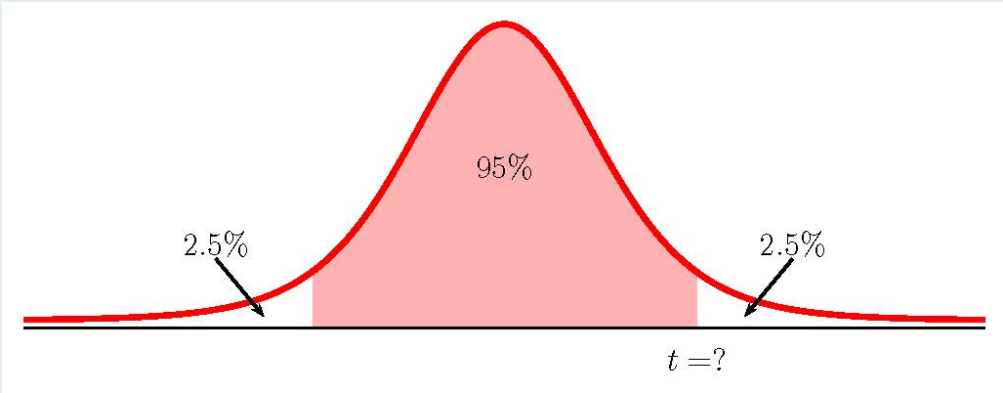
Solution

1. Let photocopier A be population 1 and photocopier B be population 2. These populations are independent because there is no relationship between the repair times for the two photocopiers. From the question, we have the following information:

Photocopier A	Photocopier B
$n_1 = 60$	$n_2 = 70$
$\bar{x}_1 = 84.2$	$\bar{x}_2 = 91.6$
$s_1 = 19.4$	$s_2 = 18.8$

To find the confidence interval, we need to find the t -score for the 95% confidence interval. This means that we need to find the t -score so that the area in the tails is $1 - 0.95 = 0.05$.

$$\begin{aligned}df &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2} \\&= \frac{\left(\frac{19.4^2}{60} + \frac{18.8^2}{70}\right)^2}{\frac{1}{60-1} \times \left(\frac{19.4^2}{60}\right)^2 + \frac{1}{70-1} \times \left(\frac{18.8^2}{70}\right)^2} \\&= 123.68 \dots \\&\Rightarrow 123\end{aligned}$$



Function	t.inv.2t
Field 1	0.05
Field 2	123
Answer	1.9794...

So $t = 1.9794 \dots$. The 95\% confidence interval is

$$\begin{aligned}
 \text{Lower Limit} &= \bar{x}_1 - \bar{x}_2 - t \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\
 &= 84.2 - 91.6 - 1.9794 \dots \times \sqrt{\frac{19.4^2}{60} + \frac{18.8^2}{70}} \\
 &= -14.06
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \bar{x}_1 - \bar{x}_2 + t \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\
 &= 84.2 - 91.6 + 1.9794 \dots \times \sqrt{\frac{19.4^2}{60} + \frac{18.8^2}{70}} \\
 &= -0.74
 \end{aligned}$$

2. We are 95% confident that the difference in the mean repair time for the two photocopiers is between -14.06 minutes and -0.74 minutes.
3. Because 0 is outside the confidence interval and both limits are negative, it suggests that the difference in the means $\mu_1 - \mu_2$ is less than 0. That is, $\mu_1 - \mu_2 < 0$ ($\mu_1 < \mu_2$). This suggests that the mean for population 1 (photocopier A) is less than the mean for population 2 (photocopier B). So the mean repair time for photocopier A is less than the mean repair time for photocopier B.

NOTES

1. When calculating the limits for the confidence interval, keep all of the decimals in the t -score and other values throughout the calculation. This will ensure that there is no round-off error in the answers. Use Excel to do the calculation of the limits, clicking on the cells containing the t -score and any other values, to ensure that all of the decimal places are used in the calculation.

2. When writing down the interpretation of the confidence interval, make sure to include the confidence level, the actual difference in the population means captured by the confidence interval (i.e. be specific to the context of the question), and appropriate units for the limits.
3. The value of the degrees of freedom must be a whole number. After using the formula, remember to round the value **down** to the next whole number to get the required degrees of freedom for the t -distribution.

Conducting a Hypothesis Test for the Difference in Two Independent Population Means with Unknown Population Standard Deviations

Follow these steps to perform a hypothesis test on the difference in two independent population means with unknown population standard deviations:

1. Write down the null hypothesis that there is no difference in the population means:

$$H_0 : \mu_1 - \mu_2 = 0$$

The null hypothesis is always the claim that the two population means are equal ($\mu_1 = \mu_2$).

2. Write down the alternative hypotheses in terms of the difference in the population means. The alternative hypothesis will be one of the following:

$$H_a : \mu_1 - \mu_2 < 0 \quad (\mu_1 < \mu_2)$$

$$H_a : \mu_1 - \mu_2 > 0 \quad (\mu_1 > \mu_2)$$

$$H_a : \mu_1 - \mu_2 \neq 0 \quad (\mu_1 \neq \mu_2)$$

3. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
4. Collect the sample information for the test and identify the significance level.
5. Assuming the population standard deviations are unknown, use a t -distribution to find the p -value (the area in the corresponding tail) for the test. The t -score and degrees of freedom are

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2}$$

6. Compare the p – value to the significance level and state the outcome of the test.

- If p – value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
- If p – value $> \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.

7. Write down a concluding sentence specific to the context of the question.

USING EXCEL TO CALCULATE THE p – value FOR A HYPOTHESIS TEST ON TWO INDEPENDENT POPULATION MEANS WITH UNKNOWN POPULATION STANDARD DEVIATIONS

Assuming that the population standard deviations are unknown, the p – value for a hypothesis test on the difference in two independent population means is the area in the tail(s) of the t –distribution.

If the p – value is the area in the left tail:

- Use the **t.dist** function to find the p – value. In the **t.dist(t-score, degrees of freedom, logic operator)** function:

- For **t-score**, enter the value of t calculated from
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

- For **degrees of freedom**, enter the degrees of freedom calculated using

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2}.$$

- For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

If the p – value is the area in the right tail:

- Use the **t.dist.rt** function to find the p – value. In the **t.dist.rt(t-score, degrees of freedom)** function:

- For **t-score**, enter the value of t calculated from
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

- For **degrees of freedom**, enter the degrees of freedom calculated using

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2}.$$

If the p – value is the sum of the area in the two tails:

- Use the **t.dist.2t** function to find the p – value. In the **t.dist.2t(t-score, degrees of freedom)** function:

- For **t-score**, enter the **absolute value** of t calculated from
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$
 Note: In the **t.dist.2t** function, the value of the t

-score must be a **positive** number. If the t -score is negative, enter the absolute value of the t -score into the **t.dist.2t** function.

- For **degrees of freedom**, enter the degrees of freedom calculated using

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2}.$$

NOTE

The degrees of freedom for a t -distribution **must** be a **whole number**. The output from the degrees

of freedom formula $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2}$ is almost never a whole

number. After calculating the value of df using the formula, **round the value down to the next whole number**. Remember to enter the rounded down value of df for the degrees of freedom in the **t.dist** functions.

EXAMPLE

A researcher wants to study the difference between the average amount of time boys and girls aged seven to eleven spend playing sports each day. In a sample of 9 girls, the average number of hours spent playing sports per day is 2 hours with a standard deviation of 0.866 hours. In a sample of 16 boys, the average number of hours spent playing sports per day is 3.2 hours with a standard deviation of 1 hours. Both populations have a normal distribution. At the 5% significance level, is there a difference in the mean amount of time boys and girls aged seven to eleven play sports each day?

Solution

Let girls be population 1 and boys be population 2. These populations are independent because there is no relationship between the two groups. From the questions, we have the following information:

Girls	Boys
$n_1 = 9$	$n_2 = 16$
$\bar{x}_1 = 2$	$\bar{x}_2 = 3.2$
$s = 0.866$	$s_2 = 1$

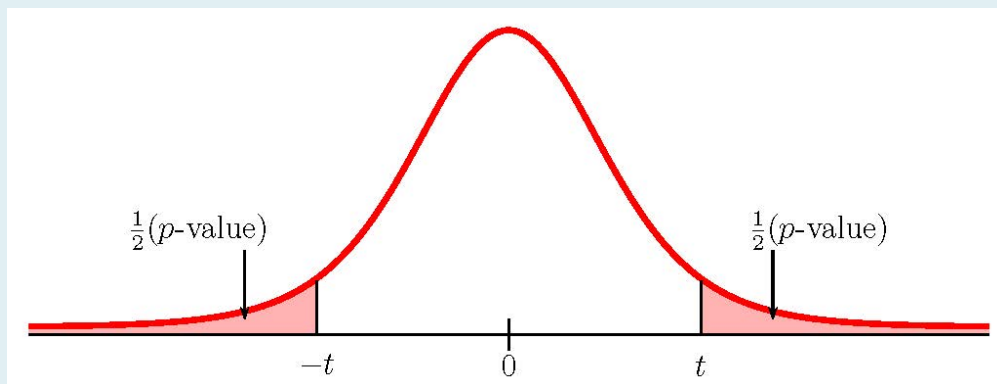
Hypotheses:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

 p – value:

This is a test on a the difference in two population means where the population standard deviation are unknown. So we use a t -distribution to calculate the p – value. Because the alternative hypothesis is a \neq , the p – value is the sum of areas in the tails of the distribution.



To use the **t.dist.2t** function, we need to calculate out the t -score and the degrees of freedom:

$$\begin{aligned}
 t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\
 &= \frac{(2 - 3.2) - 0}{\sqrt{\frac{0.866^2}{9} + \frac{1^2}{16}}} \\
 &= -3.1423\dots
 \end{aligned}$$

$$\begin{aligned}
 df &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2} \\
 &= \frac{\left(\frac{0.866^2}{9} + \frac{1^2}{16}\right)^2}{\frac{1}{9-1} \times \left(\frac{0.866^2}{9}\right)^2 + \frac{1}{16-1} \times \left(\frac{1^2}{16}\right)^2} \\
 &= 18.846\dots \\
 &\Rightarrow 18
 \end{aligned}$$

Function	t.dist.2t
Field 1	3.1423...
Field 2	18
Answer	0.0056

So the p - value = 0.0056.

Conclusion:

Because p - value = 0.0056 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that there is a difference in the mean amount of time boys and girls aged seven to eleven play sports each day.

NOTES

1. The null hypothesis $\mu_1 - \mu_2 = 0$ is the claim that there is no difference in the mean amount of time boys and girls spend playing sports each day ($\mu_1 = \mu_2$). That is, the two populations have the same mean.
2. The alternative hypothesis $\mu_1 - \mu_2 \neq 0$ is the claim that there is a difference in the mean amount of time boys and girls spend playing sports each day ($\mu_1 \neq \mu_2$). That is, the two populations have different means.
3. Keep all of the decimals throughout the calculation (i.e. in the t -score, etc.) to avoid any round-off error in the calculation of the p - value. This ensures that we get the most accurate value for the p - value. Use Excel to do the calculations, and then click on the cells in subsequent calculations.
4. The value of the degrees of freedom must be a whole number. After using the formula, remember to round the value **down** to the next whole number to get the required degrees of freedom for the t -distribution.
5. The **t.dist.2t** function requires that the value entered for the t -score is **positive**. A negative t -score entered into the **t.dist.2t** function generates an error in Excel. In this case, the value of the t -score is negative, so we must enter the absolute value of this t -score into field 1.
6. The p - value of **0.0056** is a small probability compared to the significance level, and so is unlikely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, there is a difference in the mean amount of time boys and girls spend playing sports each day.

EXAMPLE

A town has two colleges. A local community group believes that students who graduate from College A have taken more math classes than the students who graduate from College B. In a sample of 11

graduates from College A, the average is 4 math classes per graduate with a standard deviation of 1.5 math classes. In a sample of 9 graduates from College B, the average is 3.5 math classes per graduate with a standard deviation of 1 math class. Both populations have a normal distribution. At the 1% significance level test the community groups claim that graduates from College A have taken more math classes than graduates from College B.

Solution

Let College A be population 1 and College B be population 2. These populations are independent because there is no relationship between the two groups. From the questions, we have the following information:

College A	College B
$n_1 = 11$	$n_2 = 9$
$\bar{x}_1 = 4$	$\bar{x}_2 = 3.5$
$s_1 = 1.5$	$s_2 = 1$

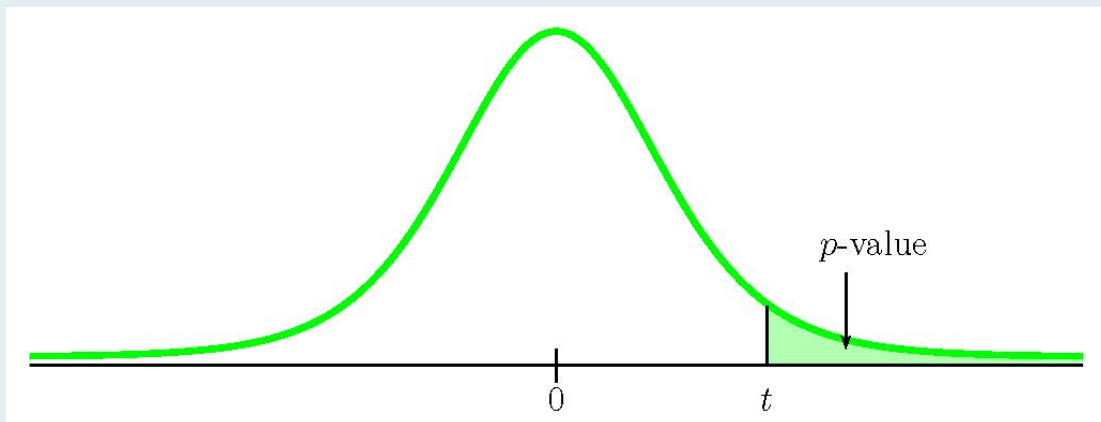
Hypotheses:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 > 0$$

p – value:

This is a test on the difference in two population means where the population standard deviation is unknown. So we use a t -distribution to calculate the p – value. Because the alternative hypothesis is a $>$, the p – value is the area in the right tail of the distribution.



To use the **t.dist.rt** function, we need to calculate out the t -score and the degrees of freedom:

$$\begin{aligned}
 t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\
 &= \frac{(4 - 3.5) - 0}{\sqrt{\frac{1.5^2}{11} + \frac{1^2}{9}}} \\
 &= 0.8899 \dots
 \end{aligned}$$

$$\begin{aligned}
 df &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2} \\
 &= \frac{\left(\frac{1.5^2}{11} + \frac{1^2}{9}\right)^2}{\frac{1}{11-1} \times \left(\frac{1.5^2}{11}\right)^2 + \frac{1}{9-1} \times \left(\frac{1^2}{9}\right)^2} \\
 &= 17.397 \dots \\
 &\Rightarrow 17
 \end{aligned}$$

Function	t.dist.rt
Field 1	0.8899...
Field 2	17
Answer	0.1930

So the p – value = 0.1930.

Conclusion:

Because p – value = 0.1930 > 0.01 = α , we do not reject the null hypothesis. At the 1\% significance level, there is not enough evidence to suggest that, on average, graduates of College A take more math classes than graduates of College B.

NOTES

1. The null hypothesis $\mu_1 - \mu_2 = 0$ is the claim that the average number of math classes taken by graduates of College A equals the average number of math classes taken by graduates of College B ($\mu_1 = \mu_2$). That is, the two populations have the same mean.
2. The alternative hypothesis $\mu_1 - \mu_2 > 0$ is the claim that, on average, graduates of College A take more math classes than graduates of College B ($\mu_1 > \mu_2$).
3. Keep all of the decimals throughout the calculation (i.e. in the t -score, etc.) to avoid any round-off error in the calculation of the ($\mu_1 = \mu_2$). This ensures that we get the most accurate value for the ($\mu_1 = \mu_2$). Use Excel to do the calculations, and then click on the cells in subsequent calculations.
4. The value of the degrees of freedom must be a whole number. After using the formula, remember to round the value **down** to the next whole number to get the required degrees of freedom for the t -distribution.
5. The p — value of **0.1930** is a large probability compared to the significance level, and so is likely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, graduates from the two colleges take, on average, the same number of math classes.

EXAMPLE

A professor at a large community college taught both an online section and a face-to-face section of his statistics course. The professor wants to study the difference in the average score on the final exam, believing that the mean score for the online section would be lower than the face-to-face section. The professor randomly selected 30 final exam scores from each section and recorded the scores in the tables below.

Online Section:

67.6	41.2	85.3	55.9	82.4	91.2	73.5	94.1	64.7	64.7
70.6	38.2	61.8	88.2	70.6	58.8	91.2	73.5	82.4	35.5
94.1	88.2	64.7	55.9	88.2	97.1	85.3	61.8	79.4	79.4

Face-to-Face Section:

77.9	95.3	81.2	74.1	98.8	88.2	85.9	92.9	87.1	88.2
69.4	57.6	69.4	67.1	97.6	85.9	88.2	91.8	78.8	71.8
98.8	61.2	92.9	90.6	97.6	100	95.3	83.5	92.9	89.4

At the 5% significance level, is the mean of the final exam score for the online section lower than the mean of the final exam score for the face-to-face section?

Solution

Let the online section be population 1 and the face-to-face section be population 2. These populations are independent because there is no relationship between the two groups. From the questions, we have the following information:

Online	Face-to-Face
$n_1 = 30$	$n_2 = 30$
$\bar{x}_1 = 72.85$	$\bar{x}_2 = 84.98$
$s_1 = 16.918 \dots$	$s_2 = 11.714 \dots$

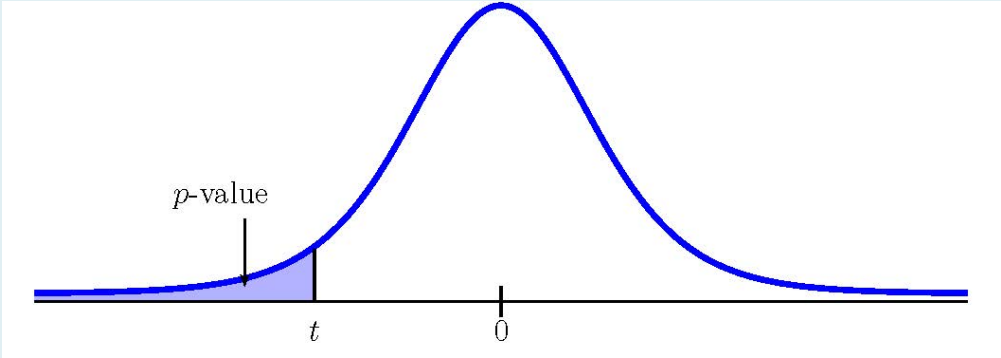
Hypotheses:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 < 0$$

 p – value:

This is a test on the difference in two population means where the population standard deviation are unknown. So we use a t -distribution to calculate the p – value. Because the alternative hypothesis is a $<$, the p – value is the area in the left tail of the distribution.



To use the **t.dist** function, we need to calculate out the t -score and the degrees of freedom:

$$\begin{aligned} t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{(72.85 - 84.98) - 0}{\sqrt{\frac{16.918\dots^2}{30} + \frac{11.714\dots^2}{30}}} \\ &= -3.228\dots \end{aligned}$$

$$\begin{aligned} df &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2} \\ &= \frac{\left(\frac{16.918\dots^2}{30} + \frac{11.714\dots^2}{30}\right)^2}{\frac{1}{30-1} \times \left(\frac{16.918\dots^2}{30}\right)^2 + \frac{1}{30-1} \times \left(\frac{11.714\dots^2}{30}\right)^2} \\ &= 51.608\dots \\ &\Rightarrow 51 \end{aligned}$$

Function	t.dist
Field 1	-3.228...
Field 2	51
Field 3	true
Answer	0.0011

So the p – value = 0.0011.

Conclusion:

Because $p - \text{value} = 0.0011 < 0.05 = \alpha$, we do reject the null hypothesis in favour of the alternative hypothesis. At the 5\% significance level, there is enough evidence to suggest that the mean final exam score for the online section is lower than the face-to-face section.

NOTES

1. The null hypothesis $\mu_1 - \mu_2 = 0$ is the claim that the average final exam score is the same for both sections ($\mu_1 = \mu_2$). That is, the two populations have the same mean.
2. The alternative hypothesis $\mu_1 - \mu_2 < 0$ is the claim that the average final exam score for the online section is lower than the face-to-face section ($\mu_1 < \mu_2$).
3. Keep all of the decimals throughout the calculation (i.e. in the sample means, sample standard deviations, in the t -score, etc.) to avoid any round-off error in the calculation of the $p - \text{value}$. This ensures that we get the most accurate value for the $p - \text{value}$. Use Excel to do the calculations, and then click on the cells in subsequent calculations.
4. The value of the degrees of freedom must be a whole number. After using the formula, remember to round the value **down** to the next whole number to get the required degrees of freedom for the t -distribution.
5. The $p - \text{value}$ of **0.0011** is a small probability compared to the significance level, and so is unlikely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the average final exam score for the online section is lower than for the face-to-face section.

TRY IT

A study is done to determine if Company A retains its workers longer than Company B. Company A samples 15 workers, and their average time with the company is 5 years with a standard deviation of

1.2 years. Company B samples 20 workers and their average time with the company is 4.5 years with a standard deviation of 0.8 years. The populations are normally distributed. At the 5% significance level, on average, do workers at Company A stay longer than workers at Company B?

Click to see Solution

Let Company A be population 1 and Company B be population 2.

Hypotheses:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 > 0$$

p – value:

$$\begin{aligned} t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{(5 - 4.5) - 0}{\sqrt{\frac{1.2^2}{15} + \frac{0.8^2}{20}}} \\ &= 1.3975 \dots \end{aligned}$$

$$\begin{aligned} df &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2} \\ &= \frac{\left(\frac{1.2^2}{15} + \frac{0.8^2}{20}\right)^2}{\frac{1}{15-1} \times \left(\frac{1.2^2}{15}\right)^2 + \frac{1}{20-1} \times \left(\frac{0.8^2}{20}\right)^2} \\ &= 23.005 \dots \\ &\Rightarrow 23 \end{aligned}$$

Function	t.dist.rt
Field 1	1.3975...
Field 2	23
Answer	0.0878

Conclusion:

Because $p - \text{value} = 0.0878 > 0.05 = \alpha$, we do not reject the null hypothesis. At the 5% significance level, there is not enough evidence to suggest that, on average, workers at Company A stay longer than workers at Company B.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=223#oembed-1>

Video: “Excel 2013 Statistical Analysis #66: Confidence Interval for Population Differences Sigma NOT Known” by excelisfun [16:12] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=223#oembed-2>

Video: “Excel 2013 Statistical Analysis #67: Hypothesis Testing for Population Differences Sigma NOT Known” by excelisfun [17:29] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A study is done to determine if students in the California State University system take longer to graduate, on average, than students enrolled in private universities. In a sample of 100 California State University system students, the mean time to graduate was 4.3 years with a standard deviation of 0.8 years. In a sample of 100 private university students, the mean time to graduate was 4.1 years with a standard deviation of 0.3 years. At the 5% significance level, do students in the California state university system take longer to graduate, on average, than students at private universities? (Note: use 126 for the degrees of freedom.)

Click to see Answer

Let California state university students be population 1 and private university students be population 2.

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$
 $H_a : \mu_1 - \mu_2 > 0$
- $p - \text{value} = 0.0104$
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the mean graduation time for students in the California State University system is longer than the mean graduation time for students at private universities.

2. It is believed that the mean grade on an English essay in a particular school system for males is lower than for females. In a random sample of 35 males, the mean score was 80 with a standard deviation of 6. In a random sample of 31 females, the mean score was 82 with a standard deviation of 3. At the 1% significance level, test if the mean grade on an English essay is lower for males than females. (Note: use 51 for the degrees of freedom.)

Click to see Answer

Let males be population 1 and females be population 2.

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$
 $H_a : \mu_1 - \mu_2 < 0$
- $p - \text{value} = 0.0322$
- Conclusion: At the 1% significance level, there is not enough evidence to conclude that the mean grade on an English essay is lower for males than females.

3. We are interested in whether children's entertainment software costs more, on average, than children's educational computer software. In a sample of 35 entertainment software titles, the mean cost was \$33.86 with a standard deviation of \$10.87. In a sample of 36 educational software titles, the mean cost was \$31.14 with a standard deviation of \$4.69. At the 5% significance level, determine if children's entertainment software costs more, on average, than children's educational software. (Note: use 45 for the degrees of freedom.)

Click to see Answer

Let entertainment software be population 1 and educational software be population 2.

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$
 $H_a : \mu_1 - \mu_2 > 0$
- p - value = 0.0792
- Conclusion: At the 5% significance level, there is not enough evidence to conclude that the mean cost of children's entertainment software is greater than the mean cost of children's educational software.

4. A student at a four-year college claims that mean enrollment at four-year colleges is the same as two-year colleges. In a sample of 35 four-year colleges, the mean enrollment was 5,789 with a standard deviation of 2,097. In a sample of 35 two-year colleges, the mean enrollment was 5,068 with a standard deviation of 1,765. At the 5% significance level, is there a difference in the mean enrollment at four-year colleges and two-year colleges? (Note: use 66 for the degrees of freedom.)

Click to see Answer

Let four-year colleges be population 1 and two-year colleges be population 2.

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$
 $H_a : \mu_1 - \mu_2 \neq 0$
- p - value = 0.0732
- Conclusion: At the 5% significance level, there is not enough evidence to conclude that there is a difference in mean enrollment at four-year colleges and two-year colleges.

5. A study is done to determine which of the two soft drinks has more sugar. In a sample of 37 cans of Beverage A, the mean amount of sugar is 36 grams with a standard deviation of 3.9 grams. In a sample of 45 cans of Beverage B, the mean amount of sugar is 38 grams with a

standard deviation of 3.1 grams. At the 5% significance level, determine if the mean amount of sugar in Beverage A is less than in Beverage B. (Note: use 68 for the degrees of freedom.)

Click to see Answer

Let Beverage A be population 1 and Beverage B be population 2.

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$
 $H_a : \mu_1 - \mu_2 < 0$
- p – value = 0.0032
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the mean amount of sugar in Beverage A is less than in Beverage B.

6. The mean number of English courses taken in a two-year time period by male and female college students is believed to be about the same. In a sample of 32 male students, the mean number of English courses was 3 with a standard deviation of 0.8. In a sample of 41 female students, the mean number of English courses was 4 with a standard deviation 1. At a 1% significance level, is there a difference in the mean number of English courses taken by male and female students? (Note: use 69 for the degrees of freedom.)

Click to see Answer

Let males be population 1 and females be population 2.

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$
 $H_a : \mu_1 - \mu_2 \neq 0$
- p – value = 0.0013
- Conclusion: At the 1% significance level, there is enough evidence to conclude that the mean number of English courses taken by male students is different from the mean number of English courses taken by female students.

7. A market company is studying the mean number of ringtones teenage girls and teenage boys use on their cell phones. In a sample of 40 randomly chosen teenage girls, the mean number of ring tones was 3.2 with a standard deviation of 1.5. In a sample of 40 randomly chosen teenage boys, the mean number of ring tones was 1.7 with a standard deviation of 0.8. At the 5% significance level, is there a difference in the mean number of ringtones for teenage girls and teenage boys? (Note: use 68 for the degrees of freedom.)

Click to see Answer

Let teenage girls be population 1 and teenage boys be population 2.

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$
 $H_a : \mu_1 - \mu_2 \neq 0$
- p - value = 0.0712
- Conclusion: At the 5\% significance level, there is not enough evidence to conclude that the mean number of ringtones for teenage girls is different than for teenage boys.

8. A dietician is studying the mean weight loss from two different diets. In a sample of 49 people on the powder diet, the mean weight loss was 42 pounds with a standard deviation of 12 pounds. In a sample of 36 people on the liquid diet, the mean weight loss was 45 pounds with a standard deviation of 14 pounds. (Note: use 68 for the degrees of freedom.)
- a. Construct a 94\% confidence interval for the difference in the mean weight loss for the powder and liquid diets.
 - b. Interpret the confidence interval in part (a).
 - c. Is it reasonable to claim that the mean weight loss with the powder diet is less than the liquid diet? Explain.

Click to see Answer

Let the powder diet be population 1, and the liquid diet be population 2.

- a. Lower Limit = -8.54 , Upper Limit = 2.54
 - b. There is a 94\% probability that the difference in the mean weight loss for the powder and liquid diets is between -8.54 pounds and 2.54 pounds.
 - c. No. Because 0 is inside the confidence interval, it suggests that the difference in the means is 0. That is $\mu_1 - \mu_2 = 0$ or $\mu_1 = \mu_2$. So, the mean weight loss for the two diets is the same.
9. A car company is studying the difference in the mean miles-per-gallon of its non-hybrid and hybrid sedan cars. In a sample of 40 hybrid sedans, the mean was 31 mpg with a standard deviation of 7 mpg. In a sample of 31 non-hybrid sedans, the mean was 22 mpg with a standard deviation of 4 mpg. (Note: use 64 for the degrees of freedom.)
- a. Construct a 95\% confidence interval for the difference in the mean miles-per-gallon in hybrid and non-hybrid cars.
 - b. Interpret the confidence interval in part (a).
 - c. Is it reasonable to claim that the mean mpg for hybrid cars is higher than for non-hybrid cars? Explain.

Click to see Answer

Let hybrid cars be population 1 and non-hybrid cars be population 2.

- a. Lower Limit = 6.36, Upper Limit = 11.64
- b. There is a 95\% probability that the difference in the mean miles-per-gallon for the hybrid and non-hybrid cars is between 6.36 mpg and 11.64 mpg.
- c. Yes. Because both limits are positive, it suggests that the difference in the means is positive. That is $\mu_1 - \mu_2 > 0$ or $\mu_1 > \mu_2$. So, the mean mpg for hybrid cars is higher than non-hybrid cars.

10. The recruiting office at a local college is studying the difference in the mean entry-level salaries for graduates with mechanical engineering degrees and electrical engineering degrees. In a sample of 50 graduates with a mechanical engineering degree, the mean entry-level salary was \$46, 100 with a standard deviation of \$3, 450. In a sample of 60 graduates with an electrical engineering degree, the mean entry-level salary was \$48, 300 with a standard deviation of \$4, 210. (Note: use 107 for the degrees of freedom.)
 - a. Construct a 99\% confidence interval for the difference in the mean entry-level salary for graduates with mechanical engineering degrees and electrical engineering degrees.
 - b. Interpret the confidence interval in part (a).
 - c. Is it reasonable to claim that the mean entry-level salary for mechanical engineering graduates is lower than for electrical engineering graduates? Explain.

Click to see Answer

Let mechanical engineering graduates be population 1 and electrical engineering graduates be population 2.

- a. Lower Limit = $-4, 115.46$, Upper Limit = -284.54
 - b. There is a 99\% probability that the difference in the mean entry-level salary for graduates with mechanical engineering degrees and electrical engineering degrees is between $-\$4, 115.46$ and $-\$284.54$.
 - c. Yes. Because both limits are negative, it suggests that the difference in the means is negative. That is $\mu_1 - \mu_2 < 0$ or $\mu_1 < \mu_2$. So, the mean entry-level salary for mechanical engineering graduates is less than for electrical engineering graduates.
-

“9.3 Statistical Inference for Two Population Means with Unknown Population Standard Deviations” and “9.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

9.3 STATISTICAL INFERENCE FOR MATCHED SAMPLES

LEARNING OBJECTIVES

- Construct and interpret a confidence interval for the mean difference for matched samples.
- Conduct and interpret hypothesis tests for matched samples.

The comparison of two population means is very common. Often, we want to find out if the two populations under study have the same mean or if there is some difference in the two population means. The approach we take when studying two population means depends on whether the samples are **independent** or **matched**.

In a **matched sample** experiment, there is some relationship between **pairs of data** in the samples. Inferences on matched samples are typically more accurate than inferences on independent samples because matched samples reduce the variability measures to only the ones within the pairs.

EXAMPLE

In a clinical trial for a new drug, patients are tested before the drug is administered, and then the same group of patients are tested after being given the drug. This is a matched sample experiment

because the same group of patients is measured before and after the administration of the drug. In this way, there are a pair of observations (a before measurement and an after measurement) for each patient.

EXAMPLE

A manufacturing company wants to know which of two different production methods allows employees to perform a task the fastest. The table below illustrates the difference in an independent sample design and a matched sample design to test the difference in the average time it takes to perform the task using the two different methods.

Independent Sample Design	Matched Sample Design
<ul style="list-style-type: none"> The company randomly selects two different groups of employees. The employees in Group 1 perform the task using Method 1, and their times are recorded. The employees in Group 2 perform the task using Method 2, and their times are recorded. 	<ul style="list-style-type: none"> The company randomly selects one group of employees. Each of the employees in the group performs the task using both methods, and their times using each method are recorded.

In the independent sample design, there is no relationship between the two groups of employees. In the matched sample design, there is one group of employees with a pair of observations (a time from Method 1 and a time from Method 2) for each employee.

In matched sample designs, we work with the **differences** in the paired observations. We combine the two samples into a single sample by calculating out the difference between each of the paired observations. Throughout this section, we will use the following notation for the sample size, mean, and standard deviation of the **differences in the paired observations**:

Symbol for:	Symbol
Population Mean of the Differences in the Paired Data	μ_D
Population Standard Deviation of the Differences in the Paired Data	σ_D
Sample Size of the Differences in the Paired Data	n_D
Sample Mean of the Differences in the Paired Data	\bar{x}_D
Sample Standard Deviation of the Differences in the Paired Data	s_D

In order to construct a confidence interval or conduct a hypothesis test on the mean of the differences in the paired data (μ_D), we need to use the distribution of the **differences** in the paired data. In such cases, we need the distribution of the **differences** in the paired data to be normal, either because the differences are assumed to be normal or because the sample size n_D is large enough ($n_D \geq 30$).

By calculating out the differences in the paired data, we combine the two samples into a single sample consisting of the differences in the paired data. We use the differences to construct the confidence interval or conduct the hypothesis test. The confidence interval on the mean difference μ_D is a confidence interval for a single population mean. Similarly, the hypothesis test on the mean difference μ_D is actually a hypothesis test on a single population mean. In this case, we will follow the exact same procedures as we learned previously for a single population mean confidence interval and hypothesis test, only now the single population consists of the differences in the paired data.

When working with a matched sample design and the differences in the paired data, the population standard deviation of the differences will be unknown. So we will need to estimate the population standard deviation with the sample standard deviation. As we have seen previously, this means we must use a t -distribution in the confidence interval and hypothesis test on the mean of the differences in the paired data.

Constructing a Confidence Interval for the Difference in Two Population Means with Matched Samples

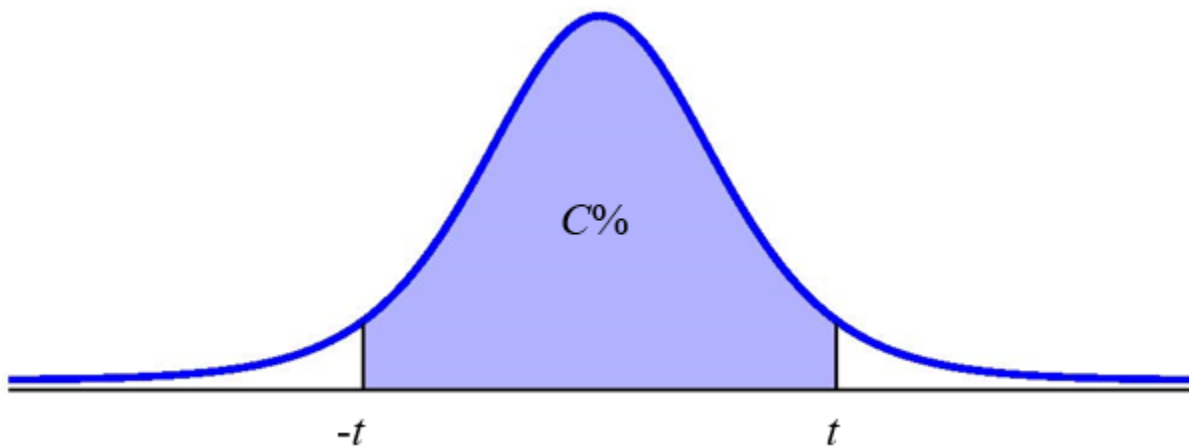
Suppose matched samples, each of size n , are taken from two related populations. The sample mean \bar{x}_D and sample standard deviation s_D for the **differences** in the matched pairs are

calculated. The limits for the confidence interval with confidence level C for the mean difference μ_D are:

$$\text{LowerLimit} = \bar{x}_D - t \times \frac{s_D}{\sqrt{n_D}}$$

$$\text{UpperLimit} = \bar{x}_D + t \times \frac{s_D}{\sqrt{n_D}}$$

where t is the positive t -score of the t -distribution with $df = n_D - 1$ so that the area under the curve in between $-t$ and t is $C\%$.



NOTES

1. In order to construct the confidence interval for the mean difference, we need to check that the distribution of the differences in the paired data follows a normal distribution. This means that we need to check that either the differences follow a normal distribution or that the sample size n_D is large enough (greater than or equal to 30).
2. When the population standard deviations are unknown, we must use a t -distribution in the construction of the confidence interval.

CALCULATING THE t -SCORE FOR A CONFIDENCE INTERVAL IN EXCEL

To find the t -score to construct a confidence interval with confidence level C , use the **t.inv.2t(area in the tails, degrees of freedom)** function.

- For **area in the tails**, enter the **sum** of the area in the tails of the t -distribution. For a confidence interval, the area in the tails is $1 - C$.
- For **degrees of freedom**, enter the degrees of freedom $df = n_D - 1$.

The output from the **t.inv.2t** function is the value of the t -score needed to construct the confidence interval.

NOTE

The **t.inv.2t** function requires that we enter the **sum** of the area in **both** tails. The area in the middle of the distribution is the confidence level C , so the sum of the area in both tails is the leftover area $1 - C$.

EXAMPLE

A company has two different methods that employees can use to complete a manufacturing task. A sample of workers is taken, and the time, in minutes, that each worker takes to complete the task

using each method is recorded. The data is shown in the table below. Assume the differences in the paired times have a normal distribution.

Worker	Method 1	Method 2
1	5.5	6.8
2	6.9	6.6
3	6.1	5.1
4	6	6.8
5	7	6.7
6	6.7	6.5
7	6.4	5.8
8	7	6.8
9	6.6	5.3
10	5.7	5.8
11	5.9	6.9
12	7	6.7
13	5.4	6.5
14	5.4	6.3
15	5.3	5

1. Construct a 98\% confidence interval for the mean difference in the time it takes the workers to complete the task.
2. Interpret the confidence interval found in part 1.
3. Is there evidence to suggest that the mean completion time for the two methods is the same? Explain.

Solution

1. We start by calculating out the differences in the paired data. We will calculate the differences as **Method 1-Method 2**.

Worker	Method 1	Method 2	Difference
1	5.5	6.8	-1.3
2	6.9	6.6	0.3
3	6.1	5.1	1
4	6	6.8	-0.8
5	7	6.7	0.3
6	6.7	6.5	0.2
7	6.4	5.8	0.6
8	7	6.8	0.2
9	6.6	5.3	1.3
10	5.7	5.8	-0.1
11	5.9	6.9	-1
12	7	6.7	0.3
13	5.4	6.5	-1.1
14	5.4	6.3	-0.9
15	5.3	5	0.3

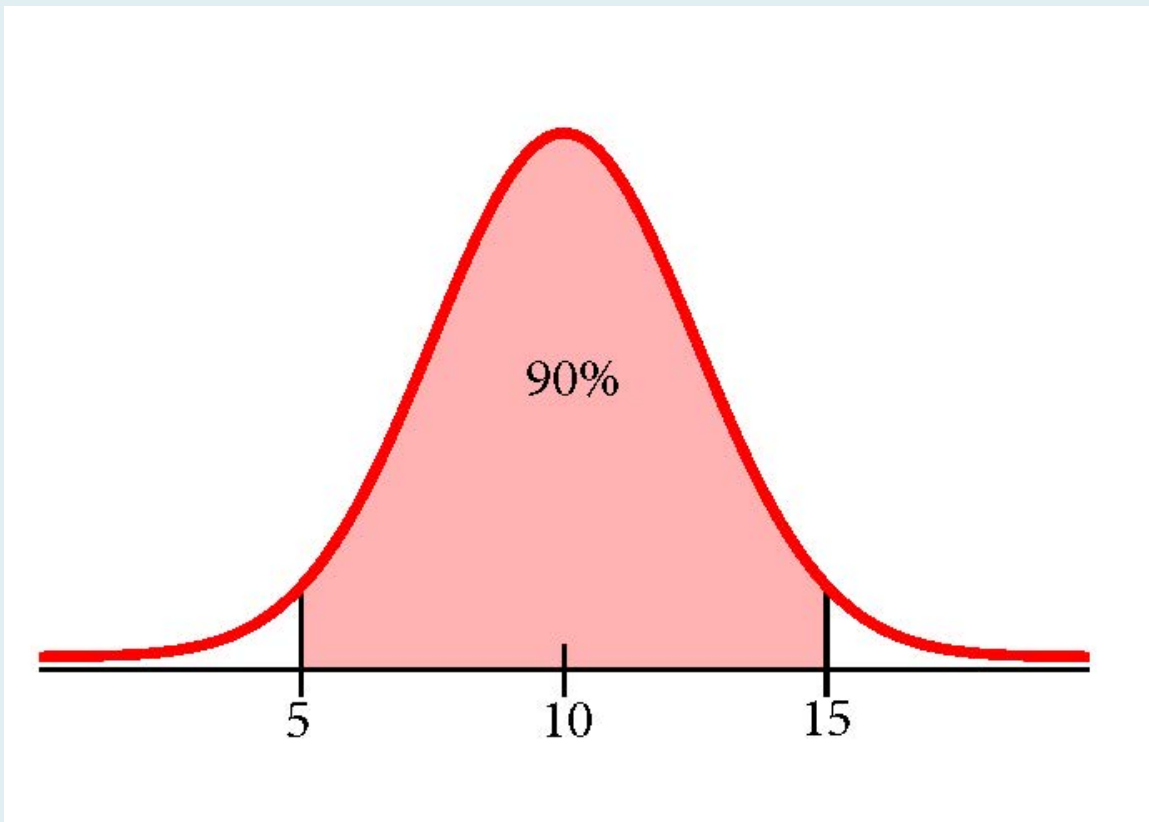
From the difference column, we have $n_D = 15$, $\bar{x}_D = -0.0466 \dots$, and $s_D = 0.7936 \dots$.

To find the confidence interval, we need to find the t -score for the 98% confidence interval.

This means that we need to find the t -score so that the sum of the area in the tails is

$1 - 0.98 = 0.02$. The degrees of freedom for the t -distribution is

$df = n_D - 1 = 15 - 1 = 14$.



Function	t.inv.2t
Field 1	0.02
Field 2	14
Answer	2.6244...

So $t = 2.6244 \dots$. The 98\% confidence interval is

$$\begin{aligned}
 \text{Lower Limit} &= \bar{x}_D - t \times \frac{s_D}{\sqrt{n_D}} \\
 &= -0.0466 \dots - 2.6244 \dots \times \frac{0.7936 \dots}{\sqrt{15}} \\
 &= -0.584
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \bar{x}_D + t \times \frac{s_D}{\sqrt{n_D}} \\
 &= -0.0466 \dots + 2.6244 \dots \times \frac{0.7936 \dots}{\sqrt{15}} \\
 &= 0.491
 \end{aligned}$$

2. We are 98\% confident that the mean difference in the completion times using the two methods is between -0.584 minutes and 0.491 minutes.
3. Because 0 is inside the confidence interval, it suggests that the mean difference μ_D is 0. That is, $\mu_D = 0$. This suggests that the mean completion times for the two methods are the same.

NOTES

1. When calculating the limits for the confidence interval, keep all of the decimals in the t -score and other values throughout the calculation. This will ensure that there is no round-off error in the answers. Use Excel to do the calculation of the differences, sample mean, sample standard deviation, and the limits, clicking on the corresponding cells to ensure that all of the decimal places are used in the calculation.
2. When writing down the interpretation of the confidence interval, make sure to include the confidence level, the actual mean difference captured by the confidence interval (i.e. be specific to the context of the question), and appropriate units for the limits.

Conducting a Hypothesis Test for the Difference in Two

Population Means with Matched Samples

Follow these steps to perform a hypothesis test on the difference in two population means with matched samples:

1. Write down the null hypothesis that the mean difference is 0:

$$H_0 : \mu_D = 0$$

The null hypothesis is always the claim that there is no difference in the two population means.

2. Write down the alternative hypotheses in terms of the mean difference. The alternative hypothesis will be **one** of the following:

$$H_a : \mu_D < 0$$

$$H_a : \mu_D > 0$$

$$H_a : \mu_D \neq 0$$

3. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
4. Collect the sample information for the test and identify the significance level.
5. Use a t -distribution to find the p – value (the area in the corresponding tail) for the test. The t -score and degrees of freedom are

$$t = \frac{\bar{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}} \quad df = n_D - 1$$

6. Compare the p – value to the significance level and state the outcome of the test.
 - If p – value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If p – value $> \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
7. Write down a concluding sentence specific to the context of the question.

USING EXCEL TO CALCULATE THE p – value FOR A HYPOTHESIS TEST ON MATCHED SAMPLES

The p – value for a hypothesis test on the mean difference of the matched samples is the area in the tail(s) of the t -distribution.

If the p – value is the area in the left tail:

- Use the **t.dist** function to find the p – value. In the **t.dist(t-score, degrees of freedom, logic operator)** function:
 - For **t-score**, enter the value of t calculated from $t = \frac{\bar{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}}$.
 - For **degrees of freedom**, enter the degrees of freedom calculated using $df = n_D - 1$.
 - For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

If the p – value is the area in the right tail:

- Use the **t.dist.rt** function to find the p – value. In the **t.dist.rt(t-score, degrees of freedom)** function:
 - For **t-score**, enter the value of t calculated from $t = \frac{\bar{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}}$.
 - For **degrees of freedom**, enter the degrees of freedom calculated using $df = n_D - 1$.

If the p – value is the sum of the area in the two tails:

- Use the **t.dist.2t** function to find the p – value. In the **t.dist.2t(t-score, degrees of freedom)** function:
 - For **t-score**, enter the **absolute value** of t calculated from $t = \frac{\bar{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}}$. Note: In the **t.dist.2t** function, the value of the t -score must be a **positive** number. If the t -score

is negative, enter the absolute value of the t -score into the **t.dist.2t** function.

- For **degrees of freedom**, enter the degrees of freedom calculated using $df = n_D - 1$

EXAMPLE

A study was conducted to investigate the effectiveness of hypnosis on reducing pain. Eight subjects are randomly selected. Each subject's pain is measured before and after being hypnotized. A lower score indicates less pain. Assume the differences in the before and after scores have a normal distribution. At the 5% significance level, are the pain sensory measurements, on average, lower after hypnotism?

Subject:	A	B	C	D	E	F	G	H
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Solution

We start by calculating out the differences in the paired data. We will calculate the differences as **before-after**.

Subject	Before	After	Difference
A	6.6	6.8	-0.2
B	6.5	2.4	4.1
C	9	7.4	1.6
D	10.3	8.5	1.8
E	11.3	8.1	3.2
F	8.1	6.1	2
G	6.3	3.4	2.9
H	11.6	2	9.6

From the difference column, we have $n_D = 8$, $\bar{x}_D = 3.125$, and $s_D = 2.911 \dots$

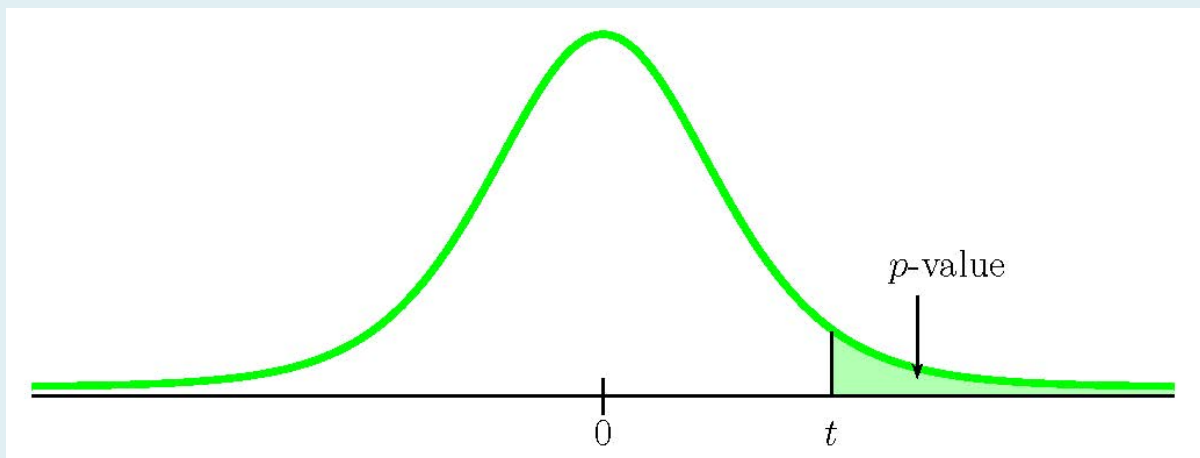
Hypotheses:

$$H_0 : \mu_D = 0$$

$$H_a : \mu_D > 0$$

p – value:

This is a test on the mean difference in matched samples, so we use a t -distribution to calculate the p – value. Because the alternative hypothesis is a $>$, the p – value is the area in the right tail of the distribution.



To use the **t.dist.rt** function, we need to calculate out the t -score:

$$\begin{aligned}
 t &= \frac{\bar{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}} \\
 &= \frac{3.125 - 0}{\frac{2.911\dots}{\sqrt{8}}} \\
 &= 3.0359\dots
 \end{aligned}$$

The degrees of freedom for the t -distribution is $n_D - 1 = 8 - 1 = 7$.

Function	t.dist.rt
Field 1	3.0359....
Field 2	7
Answer	0.0095

So the p - value = 0.0095.

Conclusion:

Because p - value = 0.0095 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level, there is enough evidence to suggest that, on average, the pain sensory measurements are lower after hypnosis.

NOTES

1. Before writing down the hypotheses, decide on the order of subtraction for calculating the differences. In a matched sample experiment, the form of the alternative hypothesis depends on the order of subtraction, so we must decide on the order of subtraction **before** writing down the hypotheses.
2. The null hypothesis $\mu_D = 0$ is the claim that there is no difference in the pain sensory measurements after hypnosis. That is, the average pain sensory measurement is the same before and after hypnosis.
3. For the alternative hypothesis, we are testing that the **after** score is lower than the **before** score. In other words, **before > after**. Because we calculated the differences as **before - after**, **before > after** means **before - after > 0**. So the alternative hypothesis is $\mu_D > 0$, the claim that the **before** score is larger than the **after** score (or the **after** score is lower than the **before** score).

4. Keep all of the decimals throughout the calculation (i.e. in the t -score, etc.) to avoid any round-off error in the calculation of the p — value. This ensures that we get the most accurate value for the p — value. Use Excel to do the calculations, and then click on the cells in subsequent calculations.
5. The p — value of 0.0095 is a small probability compared to the significance level, and so is unlikely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the after score is, on average, lower than the before score.

EXAMPLE

A study was conducted to investigate how effective a new diet was in lowering cholesterol. Nine patients were selected for the new diet and their cholesterol was measured before and after starting the new diet. The results are recorded in the table below. Assume the differences have a normal distribution. At the 5% significance level, was the new diet, on average, successful in lowering patients' cholesterol?

Subject	A	B	C	D	E	F	G	H	I
Before	209	210	205	198	216	217	238	240	222
After	199	207	189	209	217	202	211	223	201

Solution

We start by calculating out the differences in the paired data. We will calculate the differences as **after-before**.

Subject	Before	After	Difference
A	209	199	-10
B	210	207	-3
C	205	189	-16
D	198	209	11
E	216	217	1
F	217	202	-15
G	238	211	-27
H	240	223	-17
I	222	201	-21

From the difference column, we have $n_D = 9$, $\bar{x}_D = -10.777\dots$, and $s_D = 11.861\dots$

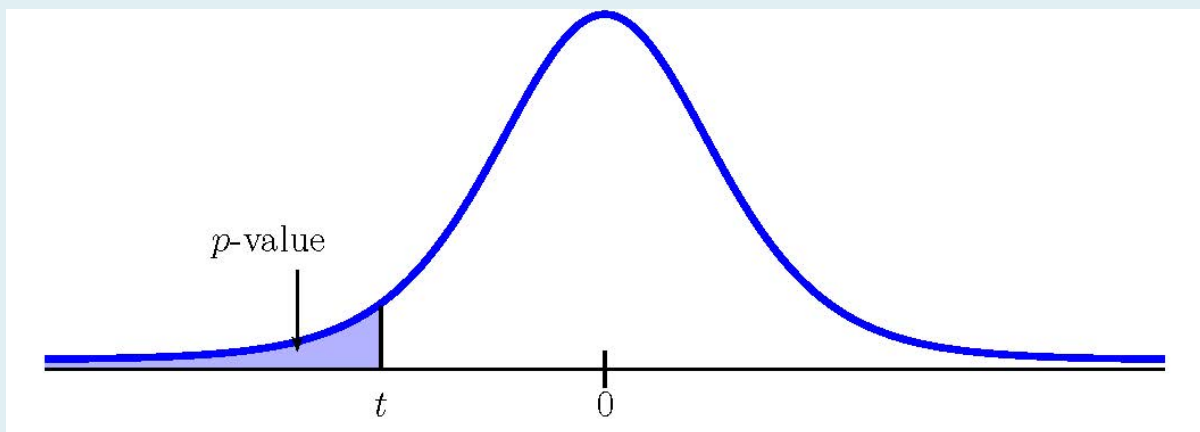
Hypotheses:

$$H_0 : \mu_D = 0$$

$$H_a : \mu_D < 0$$

p – value:

This is a test on the mean difference in matched samples, so we use a t -distribution to calculate the p – value. Because the alternative hypothesis is a $<$, the p – value is the area in the left tail of the distribution.



To use the **t.dist** function, we need to calculate out the t -score:

$$\begin{aligned}
 t &= \frac{\bar{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}} \\
 &= \frac{-10.777 \dots - 0}{\frac{11.861 \dots}{\sqrt{9}}} \\
 &= -2.725 \dots
 \end{aligned}$$

The degrees of freedom for the t -distribution is $n_D - 1 = 9 - 1 = 8$.

Function	t.dist
Field 1	-2.725...
Field 2	8
Field 3	true
Answer	0.0130

So the p - value = 0.0130.

Conclusion:

Because p - value = 0.0130 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5\% significance level, there is enough evidence to suggest that, on average, the new diet lowered the patients' cholesterol levels.

NOTES

1. Before writing down the hypotheses, decide on the order of subtraction for calculating the differences. In a matched sample experiment, the form of the alternative hypothesis depends on the order of subtraction, so we must decide on the order of subtraction **before** writing down the hypotheses.
2. The null hypothesis $\mu_D = 0$ is the claim that there is no difference in the patient's cholesterol levels. That is, the average cholesterol level is the same before and after the diet.
3. For the alternative hypothesis, we are testing that the **after** score is lower than the **before** score. In other words, **after < before**. Because we calculated the differences as **after-before**, **after < before** means **after-before < 0**, so, the alternative hypothesis is $\mu_D < 0$, the claim that the **after** score is lower than the **before** score.
4. Keep all of the decimals throughout the calculation (i.e. in the t -score, etc.) to avoid any

round-off error in the calculation of the p — value. This ensures that we get the most accurate value for the p — value. Use Excel to do the calculations, and then click on the cells in subsequent calculations.

5. The p — value of 0.0224 is a small probability compared to the significance level and so is unlikely to happen, assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the after score is, on average, lower than the before score.

EXAMPLE

Seven eighth graders at Kennedy Middle School measured how far they could push the shot-put with their dominant (writing) hand and their weaker (non-writing) hand. They thought that they could push equal distances with either hand. The results from their throws are recorded in the table below. Assume the differences are normally distributed. At the 5% significance level, is there a difference in the average distance for the dominant versus weaker hand?

Distance (in feet)	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7
Dominant Hand	30	26	34	17	19	26	20
Weaker Hand	28	14	27	18	17	26	16

Solution

We start by calculating out the differences in the paired data. We will calculate the differences as **dominant-weaker**.

Student	Dominant	Weaker	Difference
1	30	28	2
2	26	14	12
3	34	27	7
4	17	18	-1
5	19	17	2
6	26	26	0
7	20	16	4

From the difference column, we have $n_D = 7$, $\bar{x}_D = 3.714\dots$, and $s_D = 4.498\dots$

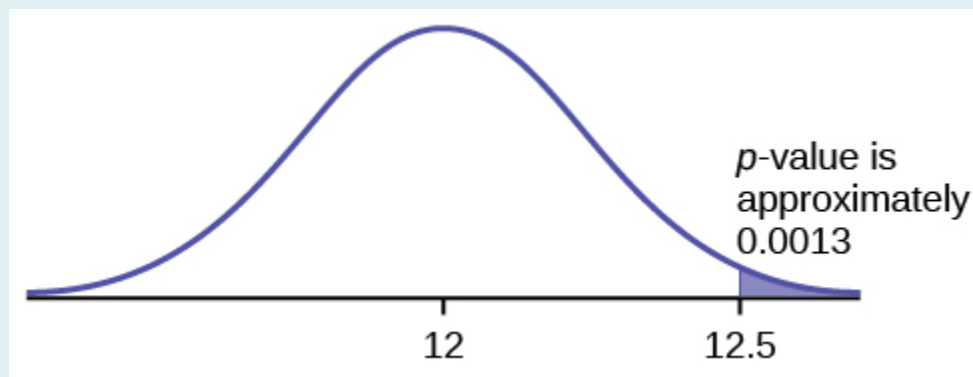
Hypotheses:

$$H_0 : \mu_D = 0$$

$$H_a : \mu_D \neq 0$$

p – value:

This is a test on the mean difference in matched samples, so we use a t -distribution to calculate the p – value. Because the alternative hypothesis is a \neq , the p – value is the sum of the area in the tails of the distribution.



To use the **t.dist.2t** function, we need to calculate out the t -score:

$$\begin{aligned}
 t &= \frac{\bar{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}} \\
 &= \frac{3.714\dots - 0}{\frac{4.498\dots}{\sqrt{7}}} \\
 &= 2.184\dots
 \end{aligned}$$

The degrees of freedom for the t -distribution is $n_D - 1 = 7 - 1 = 6$.

Function	t.dist.2t
Field 1	2.184....
Field 2	6
Answer	0.0716

So the p - value = 0.0716.

Conclusion:

Because p - value = 0.0716 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level, there is not enough evidence to suggest that there a difference in the average distance for the dominant versus weaker hand.

NOTES

1. Before writing down the hypotheses, decide on the order of subtraction for calculating the differences. In a matched sample experiment, the form of the alternative hypothesis depends on order of subtraction, so we must decide on the order of subtraction *before* writing down the hypotheses.
2. The null hypothesis $\mu_D = 0$ is the claim that there is no difference in the average distance. That is, the average distance is the same for both hands.
3. For the alternative hypothesis, we are testing that there is a difference in the dominant hand and weaker hand distances. In other words, **dominant \neq weaker**. So, the alternative hypothesis is $\mu_D \neq 0$, the claim that there is a difference in the distances.
4. Keep all of the decimals throughout the calculation (i.e. in the t -score, etc.) to avoid any round-off error in the calculation of the p - value. This ensures that we get the most accurate value for the p - value. Use Excel to do the calculations, and then click on the

cells in subsequent calculations.

5. The p — value of **0.0716** is a large probability compared to the significance level, and so is likely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, on average, the distances are the same for both hands.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=225#oembed-1>

Video: “Excel 2013 Statistical Analysis #68: Matched/Paired Samples Population Differences Sigma NOT Known” by excelisfun [20:48] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A doctor wants to know if a blood pressure medication is effective. Six subjects had their blood pressure recorded before starting the medication. After twelve weeks on the medication, the same six subjects had their blood pressure recorded again. For this test, only systolic blood pressure is of concern. Assume the differences follow a normal distribution.

Patient	Before	After
A	161	158
B	162	159
C	165	166
D	162	160
E	166	167
F	171	169

At the 1\% significance level, did the medication, on average, lower the patient's systolic blood pressure?

Click to see Answer

Calculate the differences as “before-after”.

- Hypotheses: $H_0 : \mu_D = 0$
 $H_a : \mu_D > 0$
- $p - \text{value} = 0.0699$
- Conclusion: At the 1\%significance level, there is not enough evidence to conclude that, on average, the medication lowered the patient's systolic blood pressure.

2. Ten individuals went on a low-fat diet for 12 weeks to lower their cholesterol. The data are recorded in the table below.

Individual	Starting Cholesterol Level	Ending Cholesterol Level
A	140	140
B	220	230
C	110	120
D	240	220
E	200	190
F	180	150
G	190	200
H	360	300
I	140	300
J	260	240

At the 5\% significance level, did the diet, on average, lower the individuals' cholesterol?

Click to see Answer

Calculate the differences as “starting-ending”.

- Hypotheses: $H_0 : \mu_D = 0$
 $H_a : \mu_D > 0$
- $p - \text{value} = 0.1353$
- Conclusion: At the 5\%significance level, there is not enough evidence to conclude that, on average, the diet lowered the individuals' cholesterol.

3. A local cancer support group believes that the estimate for new female breast cancer cases in the south is higher this year than last year. The group compared the estimates of new female breast cancer cases by the southern state last year and this year. The results are recorded in the table below. Assume the differences follow a normal distribution.

Southern States	Last Year	This Year
Alabama	3,450	3,720
Arkansas	2,150	2,280
Florida	15,540	15,710
Georgia	6,970	7,310
Kentucky	3,160	3,300
Louisiana	3,320	3,630
Mississippi	1,990	2,080
North Carolina	7,090	7,430
Oklahoma	2,630	2,690
South Carolina	3,570	3,580
Tennessee	4,680	5,070
Texas	15,050	14,980
Virginia	6,190	6,280

At the 1\% significance level, determine if the mean number of breast cancer cases is higher this year than last year.

Click to see Answer

Calculate the differences as “last year-this year”.

- Hypotheses: $H_0 : \mu_D = 0$
 $H_a : \mu_D < 0$
- $p - \text{value} = 0.0004$
- Conclusion: At the 1\%significance level, there is enough evidence to conclude that the mean number of breast cancer cases is higher this year than last year.

4. A traveller wanted to know if the prices of hotels are different in the ten cities that he visits the most often. The list of the cities with the corresponding hotel prices for his two favourite hotel chains is on the table. Assume the differences follow a normal distribution.

Cities	Hyatt Regency (\$)	Hilton (\$)
Atlanta	107	117
Boston	358	340
Chicago	209	219
Dallas	209	198
Denver	167	170
Indianapolis	179	185
Los Angeles	179	172
New York City	625	615
Philadelphia	179	165
Washington, DC	245	239

At the 5\% significance level, is there a difference in the mean price of hotels in the ten cities?

Click to see Answer

Calculate the differences as “Hyatt-Hilton”.

- Hypotheses: $H_0 : \mu_D = 0$
 $H_a : \mu_D \neq 0$
- $p - \text{value} = 0.2804$
- Conclusion: At the 5\%significance level, there is enough evidence to conclude that there is no difference in the mean price of hotels in the ten cities.

5. One of the questions in a study of marital satisfaction of dual-career couples was to rate the statement, “I’m pleased with the way we divide the responsibilities for childcare.” The ratings went from one (strongly disagree) to five (strongly agree). The table below contains the responses of ten couples. Assume the differences follow a normal distribution.

Couple	Wife's Score	Husband's Score
A	2	2
B	2	2
C	3	1
D	3	3
E	4	2
F	2	1
G	1	1
H	1	1
I	2	2
J	4	4

At the 5\% significance level, test if, on average, the husband's satisfaction score is higher than the wife's satisfaction score.

Click to see Answer

Calculate the differences as “wife's score-husband's score”.

- Hypotheses: $H_0 : \mu_D = 0$
 $H_a : \mu_D < 0$
- p – value = 0.0479
- Conclusion: At the 5\%significance level, there is enough evidence to conclude that, on average, the husband's satisfaction score is higher than the wife's satisfaction score.

6. A company has plans to roll out a new product on the market. Currently, the company has two different prototypes of the product but plans to only push one version onto the market. The company selects a sample of ten customers and sends both prototypes to each customer for feedback and evaluation. In particular, the company asks each customer the maximum price they would be willing to pay for each prototype. The results are shown in the table below.

Customer	Version A	Version B
A	14	17
B	16	20
C	14	17
D	16	20
E	20	18
F	15	19
G	16	20
H	13	16
I	19	15
J	17	20

At the 5\% significance level, is there a difference in the mean price customers are willing to pay for the prototypes?

Click to see Answer

Calculate the differences as “version A-version B”.

- Hypotheses: $H_0 : \mu_D = 0$
 $H_a : \mu_D \neq 0$
- $p - \text{value} = 0.0358$
- Conclusion: At the 5\%significance level, there is enough evidence to conclude that there is a difference in the mean price customers are willing to pay for the prototypes.

7. A study was conducted to test the effectiveness of a software patch in reducing system failures over a six-month period. Results for randomly selected installations are shown in the table below, recording the number of system failures before the patch was installed and the number of system failures after the patch was installed. Assume the differences have a normal distribution.

Installation	Before	After
A	3	1
B	6	5
C	4	2
D	2	0
E	5	1
F	8	0
G	2	2
H	6	2

- Construct a 97% confidence interval for the mean difference in the number of failures before and after the software patch was installed.
- Interpret the confidence interval in part (a).
- Is it reasonable to claim that the mean number of failures did not change after the software patch was installed? Explain.

Click to see Answer

Calculate the differences as “before-after”.

- Lower Limit = 0.4997 Upper Limit = 5.250
 - There is a 97% probability that the mean difference in the number of failures before and after the software patch was installed is between 0.4997 failures and 5.250 failures.
 - No. Because both limits are positive, it suggests that the mean difference μ_D is positive. That is, $\mu_D > 0$. So, before – after > 0 or before $>$ after, which suggests that the mean number of failures was smaller after the software patch was installed.
8. A study was conducted to test the effectiveness of a juggling class. Before the class started, six subjects juggled as many balls as they could at once. After the class, the same six subjects juggled as many balls as they could. Assume the differences have a normal distribution.

Juggler	Before	After
A	3	4
B	4	5
C	3	6
D	2	4
E	4	5
F	5	7

- Construct a 99\% confidence interval for the mean difference in the number of balls a subject can juggle before and after the class.
- Interpret the confidence interval in part (a).
- Is it reasonable to claim that the average mean of balls a subject can juggle is higher after the class? Explain.

Click to see Answer

Calculate the differences as “before-after”.

- Lower Limit = -3.01 Upper Limit = -0.32
- There is a 99\% probability that the mean difference in the number of balls a subject can juggle before and after the class is between -3.01 balls and -0.32 balls.
- Yes. Because both limits are negative, it suggests that the mean difference μ_D is negative. That is, $\mu_D < 0$. So, before $-$ after < 0 or before $<$ after, which suggests that the mean number of balls a subject can juggle was higher after the class.

- A company wants to assess the overall effectiveness of the sales training course on the company’s sales department. The company takes a sample of salespersons and records the number of transactions completed in a week before and after each salesperson completes the training course. Assume the differences have a normal distribution.

Salesperson	Before	After
A	20	20
B	21	24
C	23	21
D	26	23
E	22	24
F	23	21
G	20	23
H	19	24
I	22	20
J	15	21
K	15	24
L	23	19

- Construct a 93% confidence interval for the mean difference in the number of transactions completed in a week before and after training.
- Interpret the confidence interval in part (a).
- Is it reasonable to claim that the training had no effect on sales? Explain.

Click to see Answer

Calculate the differences as “before-after”.

- Lower Limit = -3.61 Upper Limit = 1.11
- There is a 93% probability that the mean difference in the number of transactions completed in a week before and after training is between -3.61 transactions and 1.11 transactions.
- Yes. Because 0 is inside the interval, it suggests that the mean difference $\mu_D = 0$. That is, $\mu_D = 0$. So, before $-$ after $= 0$ or before $=$ after, which suggests that the mean number of transactions is the same before and after training.

by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

9.4 STATISTICAL INFERENCE FOR TWO POPULATION PROPORTIONS

LEARNING OBJECTIVES

- Construct and interpret a confidence interval for two population proportions.
- Conduct and interpret hypothesis tests for two population proportions.

Similar to comparing two population means, the comparison of two population proportions is very common. Often, we want to find out if the two populations under study have the same proportion or if there is some difference in the two population proportions. Unlike two population means, we can only approach the comparison of two population proportions using independent samples. Recall that two populations are **independent** if the sample taken from population 1 is not related in any way to the sample taken from population 2. In this situation, any relationship between the samples or populations is entirely coincidental.

Throughout this section, we will use subscripts to identify the values for the proportions and sample sizes for the two populations:

Symbol for:	Population 1	Population 2
Population Proportion	p_1	p_2
Sample Size	n_1	n_2
Sample Proportion	\hat{p}_1	\hat{p}_2
Number of Items in Sample with Characteristic of Interest	x_1	x_2

In order to construct a confidence interval or conduct a hypothesis test on the difference in two

population proportions ($p_1 - p_2$), we need to use the distribution of the difference in the sample proportions $\hat{p}_1 - \hat{p}_2$.

- The mean of the distribution of the **difference** in the sample proportions is

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2.$$

- The standard deviation of the distribution of the **difference** in the sample proportions is

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 \times (1 - p_1)}{n_1} + \frac{p_2 \times (1 - p_2)}{n_2}}.$$

- The distribution of the **difference** in the sample proportions is normal if $n_1 \times p_1 \geq 5$, $n_1 \times (1 - p_1) \geq 5$, $n_2 \times p_2 \geq 5$ and $n_2 \times (1 - p_2) \geq 5$.
- Assuming the distribution of the **difference** of the sample proportions is normal, the z -score

$$\text{is } z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 \times (1 - p_1)}{n_1} + \frac{p_2 \times (1 - p_2)}{n_2}}}.$$

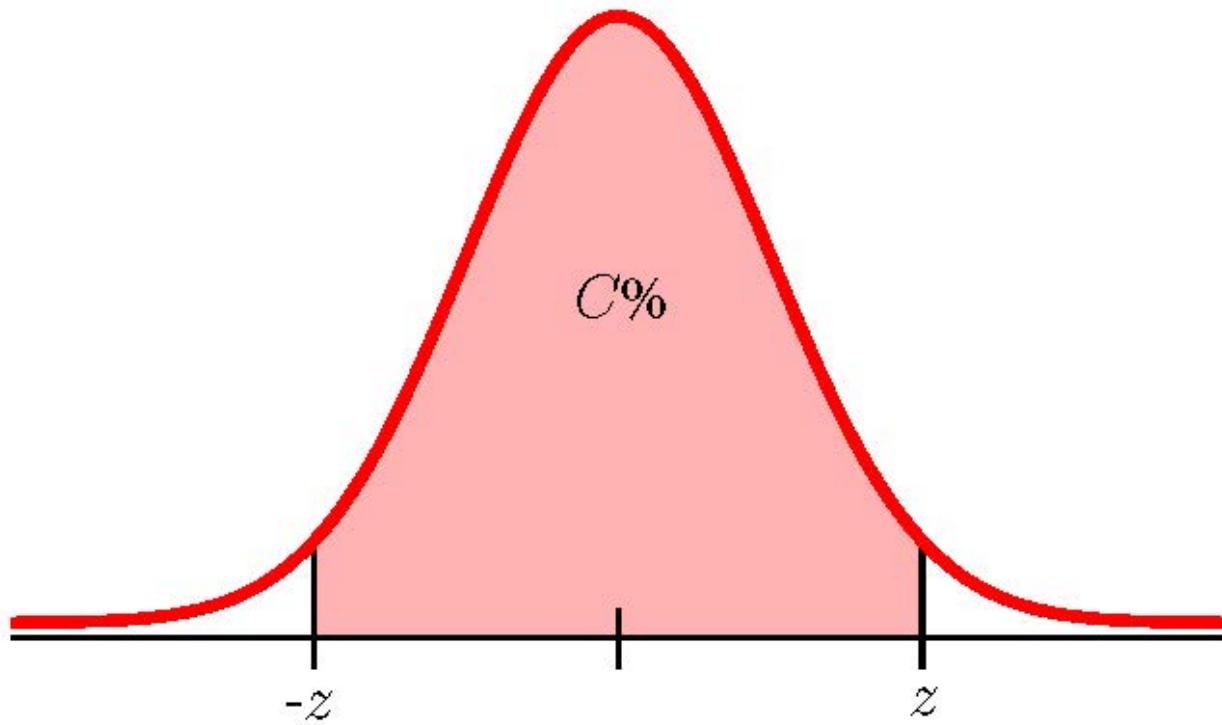
Constructing a Confidence Interval for the Difference in Two Population Proportions

Suppose a sample of size n_1 with sample proportion \hat{p}_1 is taken from population 1 and a sample of size n_2 with sample proportion \hat{p}_2 is taken from population 2. The limits for the confidence interval with confidence level C for the difference in the population proportions $p_1 - p_2$ are

$$\text{Lower Limit} = \hat{p}_1 - \hat{p}_2 - z \times \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}}$$

$$\text{Upper Limit} = \hat{p}_1 - \hat{p}_2 + z \times \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}}$$

where z is the z -score of the standard normal distribution so that the area to the left of z is $C + \frac{1 - C}{2}$.



NOTES

1. In order to construct the confidence interval for the difference in two population proportions, we need to check that the normal distribution applies. This means that we need to check that $n_1 \times p_1 \geq 5$, $n_1 \times (1 - p_1) \geq 5$, $n_2 \times p_2 \geq 5$ and $n_2 \times (1 - p_2) \geq 5$.
2. Because the population proportions p_1 and p_2 are often unknown, we replace the values of the population proportions with the sample proportions \hat{p}_1 and \hat{p}_2 in the normal distribution check. That is, when the population proportions are unknown, we check $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1 - \hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$ and $n_2 \times (1 - \hat{p}_2) \geq 5$ to verify that the normal distribution applies.

CALCULATING THE z -SCORE FOR A CONFIDENCE INTERVAL IN EXCEL

To find the z -score to construct a confidence interval with confidence level C , use the **norm.s.inv(area to the left of z)** function.

- For **area to the left of z**, enter the **entire** area to the left of the z -score you are trying to find.

For a confidence interval, the area to the left of z is $C + \frac{1 - C}{2}$.

The output from the **norm.s.inv** function is the value of the z -score needed to construct the confidence interval.

NOTE

The **norm.s.inv** function requires that we enter the **entire** area to the **left** of the unknown z -score. This area includes the confidence level C (the area in the middle of the distribution) plus the remaining area in the left tail $\frac{1 - C}{2}$.

EXAMPLE

A marketing company places an advertisement for a new brand of deodorant on two different platforms: television and social media. The company wants to study the proportion of people who remembered seeing the advertisement two hours later. In a sample of 200 people who saw the

advertisement on television, 74 remembered seeing it two hours later. In a sample of 300 people who saw the advertisement on social media, 129 remembered seeing it two hours later.

1. Construct a 98% confidence interval for the difference in the proportion of people from the two different platforms that remember seeing the advertisement two hours later.
2. Interpret the confidence interval found in part 1.
3. Is there evidence to suggest that the proportion of people from social media who remember seeing the advertisement two hours later is greater than the proportion of people from television? Explain.

Solution

1. Let television be population 1 and social media be population 2. From the question, we have the following information:

Television	Social Media
$n_1 = 200$	$n_2 = 300$
$\hat{p}_1 = \frac{74}{200} = 0.37$	$\hat{p}_2 = \frac{129}{300} = 0.43$

Before constructing the confidence interval, we check that the normal distribution applies:

$$n_1 \times \hat{p}_1 = 200 \times 0.37 = 74 \geq 5$$

$$n_1 \times (1 - \hat{p}_1) = 200 \times (1 - 0.37) = 126 \geq 5$$

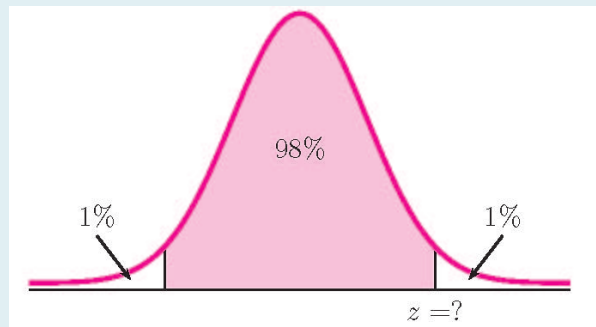
$$n_2 \times \hat{p}_2 = 300 \times 0.43 = 129 \geq 5$$

$$n_2 \times (1 - \hat{p}_1) = 300 \times (1 - 0.37) = 171 \geq 5$$

To find the confidence interval, we need to find the z -score for the 98% confidence interval.

This means that we need to find the z -score so that the entire area to the left of z is

$$0.98 + \frac{1 - 0.98}{2} = 0.99.$$



Function	norm.s.inv
Field 1	0.99
Answer	2.3263...

So $z = 2.3263 \dots$. The 98% confidence interval is

$$\begin{aligned}
 \text{Lower Limit} &= \hat{p}_1 - \hat{p}_2 - z \times \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}} \\
 &= 0.37 - 0.43 - 2.3263 \dots \times \sqrt{\frac{0.37 \times (1 - 0.37)}{200} + \frac{0.43 \times (1 - 0.43)}{300}} \\
 &= -0.1636
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \hat{p}_1 - \hat{p}_2 + z \times \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}} \\
 &= 0.37 - 0.43 + 2.3263 \dots \times \sqrt{\frac{0.37 \times (1 - 0.37)}{200} + \frac{0.43 \times (1 - 0.43)}{300}} \\
 &= 0.0436
 \end{aligned}$$

2. We are 98% confident that the difference in the proportion of people from the two platforms that remember seeing the advertisement two hours later is between -16.36% and 4.36%.
3. Because 0 is inside the confidence interval, it suggests that the difference in the proportions $p_1 - p_2$ is 0. That is, $p_1 - p_2 = 0$. This suggests that the two proportions are equal. So the proportion of people from social media who remember seeing the advertisement two hours is not greater than the proportion of people from television.

NOTES

1. Because the population proportions are unknown, we use the sample proportions in the check for normality.
2. When calculating the limits for the confidence interval, keep all of the decimals in the z -score and other values throughout the calculation. This will ensure that there is no round-off error in the answers. Use Excel to do the calculation of the limits, clicking on the cell containing the z -score and any other values, to ensure that all of the decimal places are used in the calculation.
3. The limits for the confidence interval are percents. For example, the upper limit of 0.0436 is the decimal form of a percent: 4.36\%.
4. When writing down the interpretation of the confidence interval, make sure to include the confidence level, the actual difference in the population proportions captured by the confidence interval (i.e. be specific to the context of the question), and express the limits as percents.

Conducting a Hypothesis Test for the Difference in Two Population Proportions

Follow these steps to perform a hypothesis test on the difference in two population proportions:

1. Write down the null hypothesis that there is no difference in the population proportions:

$$H_0 : p_1 - p_2 = 0$$

The null hypothesis is always the claim that the two population proportions are equal ($p_1 = p_2$).

2. Write down the alternative hypotheses in terms of the difference in the population proportions. The alternative hypothesis will be **one** of the following:

$$H_a : p_1 - p_2 < 0 \quad (p_1 < p_2)$$

$$H_a : p_1 - p_2 > 0 \quad (p_1 > p_2)$$

$$H_a : p_1 - p_2 \neq 0 \quad (p_1 \neq p_2)$$

3. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
4. Collect the sample information for the test and identify the significance level.
5. Check the conditions $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1 - \hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$ and $n_2 \times (1 - \hat{p}_2) \geq 5$ to verify that the normal distribution applies. Use the normal distribution to find the p - value (the area in the corresponding tail) for the test. The z -score is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p} \times (1 - \bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

6. Compare the p - value to the significance level and state the outcome of the test.
 - If p - value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If p - value $> \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
7. Write down a concluding sentence specific to the context of the question.

NOTES

1. Because the population proportions p_1 and p_2 are often unknown, we replace the values of the population proportions with the sample proportions \hat{p}_1 and \hat{p}_2 in the normal distribution check. That is, when the population proportions are unknown, we check $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1 - \hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$ and $n_2 \times (1 - \hat{p}_2) \geq 5$ to verify

that the normal distribution applies to the calculation of the p -value.

2. Because we are testing the equality of the two population proportions, the z -score for the hypothesis test uses a **pooled sample proportion** \bar{p} . The pooled sample proportion \bar{p} combines the sample data to create an estimate of the overall proportion of success.

USING EXCEL TO CALCULATE THE p – value FOR A HYPOTHESIS TEST ON TWO INDEPENDENT POPULATION PROPORTIONS

The p – value for a hypothesis test on the difference in two population proportions is the area in the tail(s) of the normal distribution, assuming that the conditions for using a normal distribution are met ($n_1 \times p_1 \geq 5$, $n_1 \times (1 - p_1) \geq 5$, $n_2 \times p_2 \geq 5$ and $n_2 \times (1 - p_2) \geq 5$).

The p – value is the area in the tail(s) of a normal distribution, so the **norm.dist(x,μ,σ,logic operator)** function can be used to calculate the p – value.

- For **x**, enter the value for $\hat{p}_1 - \hat{p}_2$.
- For **μ**, enter 0, the value of $p_1 - p_2$ from the null hypothesis. This is the mean of the distribution of the differences in the sample proportions.
- For **σ**, enter the value of $\sqrt{\bar{p} \times (1 - \bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$ where $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$. The value for **σ** is the bottom part of the z -score used in the hypothesis test.
- For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

As with the previous chapter, use the appropriate technique with the **norm.dist** function to find the area in the left-tail, the area in the right-tail or the sum of the area in tails.

EXAMPLE

A cell phone company claimed that iPhones are more popular with adults 30 years old or younger than with adults over 30 years old. A consumer advocacy group wants to test this claim. In a sample of 1340 adults 30 years old or younger, 134 own an iPhone. In a sample of 250 adults over the age of 30, 15 own an iPhone. At the 5% significance level, is the proportion of adults 30 years old or younger who own an iPhone greater than the proportion of adults over the age of 30 who own an iPhone?

Solution

Let adults 30 years old or younger be population 1 and adults over 30 years old be population 2. From the question, we have the following information:

30 Years or Younger	Over 30 Years
$n_1 = 1340$	$n_2 = 250$
$x_1 = 134$	$x_2 = 15$
$\hat{p}_1 = \frac{134}{1340} = 0.1$	$\hat{p}_2 = \frac{15}{250} = 0.05$

Hypotheses:

$$H_0 : p_1 - p_2 = 0$$

$$H_a : p_1 - p_2 > 0$$

p – value:

Before calculating the p – value, we check that the normal distribution applies:

$$n_1 \times \hat{p}_1 = 1340 \times 0.1 = 134 \geq 5$$

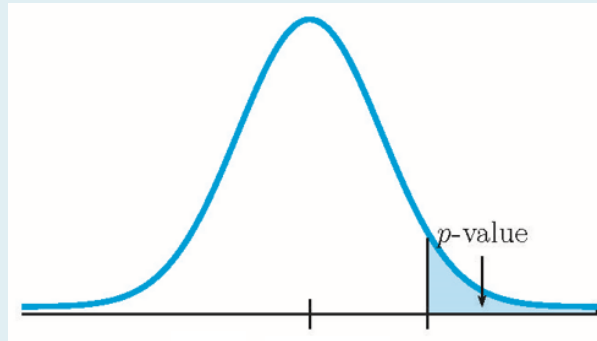
$$n_1 \times (1 - \hat{p}_1) = 1340 \times (1 - 0.1) = 1206 \geq 5$$

$$n_2 \times \hat{p}_2 = 250 \times 0.05 = 15 \geq 5$$

$$n_2 \times (1 - \hat{p}_2) = 250 \times (1 - 0.05) = 235 \geq 5$$

Because $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1 - \hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$, and $n_2 \times (1 - \hat{p}_2) \geq 5$, the normal distribution applies, and so we use a normal distribution to calculate the p – value.

Because the alternative hypothesis is a $>$, the p – value is the area in the right tail of the distribution.



The pooled sample proportion is:

$$\begin{aligned}\bar{p} &= \frac{x_1 + x_2}{n_1 + n_2} \\ &= \frac{134 + 15}{1340 + 250} \\ &= \frac{149}{1590} \\ &= 0.09371 \dots\end{aligned}$$

Function	1-norm.dist
Field 1	0.1-0.05
Field 2	0
Field 3	sqrt(0.09371... *(1-0.09371...)*(1/1340+1/250))
Field 4	true
Answer	0.0232

So the p – value = 0.0232.

Conclusion:

Because p – value = 0.0232 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that the proportion of adults 30 years old or younger who own an iPhone is greater than the proportion of adults over the age of 30 who own an iPhone.

NOTES

1. The null hypothesis $p_1 - p_2 = 0$ is the claim that the proportion of adults 30 or younger with an iPhone equals the proportion of adults over 30 with an iPhone ($p_1 = p_2$). That is, the two populations have the same proportion.
2. The alternative hypothesis $p_1 - p_2 > 0$ is the claim that the proportion of adults 30 or younger with an iPhone is greater than the proportion of adults over 30 with an iPhone ($p_1 > p_2$).
3. Make sure to keep all of the decimal places throughout the calculation to avoid any round-off error in the **p - value**. Perform the calculations of the sample proportions and the pooled sample proportion \bar{p} in Excel and then click on the corresponding cells when completing the fields in the **norm.dist** function.
4. The **p - value** is the area in the right tail of the normal distribution. In the calculation of the **p - value**:
 - The function is **1-norm.dist** because we are finding the area in the right tail of a normal distribution.
 - Field 1 is the value of $\hat{p}_1 - \hat{p}_2 = 0.1 - 0.05$.
 - Field 2 is **0**, the value of $p_1 - p_2$ from the null hypothesis. Remember, we run the test assuming the null hypothesis is true, so that means we assume $p_1 - p_2 = 0$.
 - Field 3 is the value of

$$\sqrt{\bar{p} \times (1 - \bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{0.09371 \dots \times (1 - 0.09371 \dots) \times \left(\frac{1}{1340} + \frac{1}{250} \right)}$$
5. The **p - value** of **0.0232** is a small probability compared to the significance level, and so is unlikely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the proportion of adults 30 years old or younger who own an iPhone is greater than the proportion of adults over the age of 30 who own an iPhone.

EXAMPLE

Two types of medication for hives are tested to determine if there is a difference in the proportions of adult patient reactions. In a sample of 200 adults given medication A, 20 still had hives 30 minutes after taking the medication. In a sample of 200 adults given medication B, 12 still had hives 30 minutes after taking the medication. At the 1% significance level, is there a difference in the proportion of adults who still have hives 30 minutes after taking medications?

Solution

Let medication A be population 1 and medication B be population 2. From the question, we have the following information:

Medication A	Medication B
$n_1 = 200$	$n_2 = 200$
$x_1 = 20$	$x_2 = 12$
$\hat{p}_1 = \frac{20}{200} = 0.1$	$\hat{p}_2 = \frac{12}{200} = 0.06$

Hypotheses:

$$H_0 : p_1 - p_2 = 0$$

$$H_a : p_1 - p_2 \neq 0$$

p – value:

Before calculating the p – value, we check that the normal distribution applies:

$$n_1 \times \hat{p}_1 = 200 \times 0.1 = 20 \geq 5$$

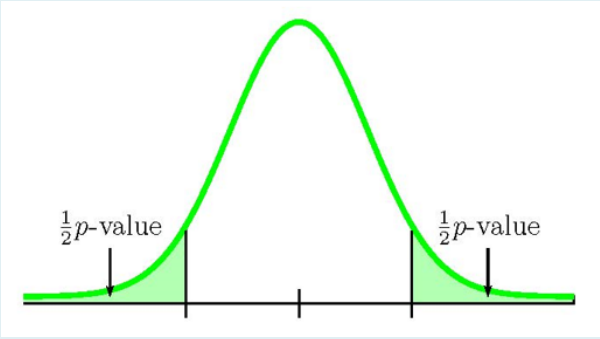
$$n_1 \times (1 - \hat{p}_1) = 200 \times (1 - 0.1) = 180 \geq 5$$

$$n_2 \times \hat{p}_2 = 200 \times 0.06 = 12 \geq 5$$

$$n_2 \times (1 - \hat{p}_2) = 200 \times (1 - 0.06) = 188 \geq 5$$

Because $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1 - \hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$, and $n_2 \times (1 - \hat{p}_2) \geq 5$, the normal distribution applies, and so we use a normal distribution to calculate the p – value.

. Because the alternative hypothesis is a \neq , the p – value is the sum of the area in the two tails of the distribution.



We need to know if the sample information relates to the left or right tail because that will determine how we calculate out the area of that tail using the normal distribution. In this case, $\hat{p}_1 > \hat{p}_2$ ($0.1 > 0.06$), so the sample information relates to the right tail of the normal distribution. This means that we will calculate out the area in the right tail using **1-norm.dist**. However, this is a two-tailed test where the *p* – value is the sum of the area in the two tails, and the area in the right tail is only one-half of the *p* – value. The area in the right tail equals the area in the left tail, and the *p* – value is the sum of these two areas.

The pooled sample proportion is:

$$\begin{aligned}\bar{p} &= \frac{x_1 + x_2}{n_1 + n_2} \\ &= \frac{20 + 12}{200 + 200} \\ &= \frac{32}{400} \\ &= 0.08\end{aligned}$$

Function	1-norm.dist
Field 1	0.1-0.06
Field 2	0
Field 3	sqrt(0.08*(1-0.08)*(1/200+1/200))
Field 4	true
Answer	0.0702

So the area in the right tail is 0.0702, which means $\frac{1}{2}p\text{ – value} = 0.0702$. This is also the area in the left tail, so

$$p - \text{value} = 0.0702 + 0.0702 = 0.1404$$

Conclusion:

Because $p - \text{value} = 0.1404 > 0.01 = \alpha$, we do not reject the null hypothesis. At the 1\% significance level, there is not enough evidence to suggest that there is a difference in the proportion of adults who still have hives 30 minutes after taking medication.

NOTES

1. The null hypothesis $p_1 - p_2 = 0$ is the claim that there is no difference in the proportion of adults with hives 30 minutes after taking the medications ($p_1 = p_2$). That is, the two populations have the same proportion.
2. The alternative hypothesis $p_1 - p_2 \neq 0$ is the claim that there is a difference in the proportion of adults with hives 30 minutes after taking the medications ($p_1 \neq p_2$).
3. Make sure to keep all of the decimal places throughout the calculation to avoid any round-off error in the $p - \text{value}$. Perform the calculations of the sample proportions and the pooled sample proportion \bar{p} in Excel and then click on the corresponding cells when completing the fields in the **norm.dist** function.
4. In a two-tailed hypothesis test that uses the normal distribution, we will only have sample information relating to **one** of the two tails. We must determine which of the tails the sample information belongs to and then calculate out the area in that tail. The area in each tail represents exactly half of the $p - \text{value}$, so the $p - \text{value}$ is the sum of the areas in the two tails.

- If the sample proportion \hat{p}_1 is less than the sample proportion \hat{p}_2 ($\hat{p}_1 < \hat{p}_2$), the sample information belongs to the **left tail**.

- We use **norm.dist** $(\hat{p}_1 - \hat{p}_2, 0, \sqrt{\bar{p} \times (1 - \bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \text{true})$ to find the area in the left tail. The area in the right tail equals the area in the left tail, so we can find the $p - \text{value}$ by adding the output from this function to itself.

- If the sample proportion \hat{p}_1 is greater than the sample proportion \hat{p}_2 ($\hat{p}_1 > \hat{p}_2$), the sample information belongs to the **right tail**.

- We use **1-norm.dist**($\hat{p}_1 - \hat{p}_2, 0, \sqrt{\bar{p} \times (1 - \bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, **true**) to find the area in the right tail. The area in the left tail equals the area in the right tail, so we can find the p — **value** by adding the output from this function to itself.

5. The p — **value** of **0.1404** is a large probability compared to the significance level, and so is likely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, there is no difference in the proportion of adults with hives 30 minutes after taking the medications.

EXAMPLE

A valve manufacturer recently launched a new valve, Valve A, and they want to claim that the proportion of their valves that fail under 4500 psi is the smallest of all the other valves on the market. The manufacturer decides to compare Valve A with the most popular valve on the market, Valve B. In a sample of **100** Valve A's, **6** failed at 4500 psi. In a sample of **150** Valve B's, **16** failed at 4500 psi. At the 5% significance level, is the proportion of Valve As that fail under 4500 psi less than the proportion of Valve Bs that fail under 4500 psi?

Solution

Let Valve A be population 1 and Valve B be population 2. From the question, we have the following information:

Valve A	Valve B
$n_1 = 100$	$n_2 = 150$
$x_1 = 6$	$x_2 = 16$
$\hat{p}_1 = \frac{6}{100} = 0.06$	$\hat{p}_2 = \frac{16}{150} = 0.1066\dots$

Hypotheses:

$$H_0 : p_1 - p_2 = 0$$

$$H_a : p_1 - p_2 < 0$$

p – value:

Before calculating the p – value, we check that the normal distribution applies:

$$n_1 \times \hat{p}_1 = 100 \times 0.06 = 6 \geq 5$$

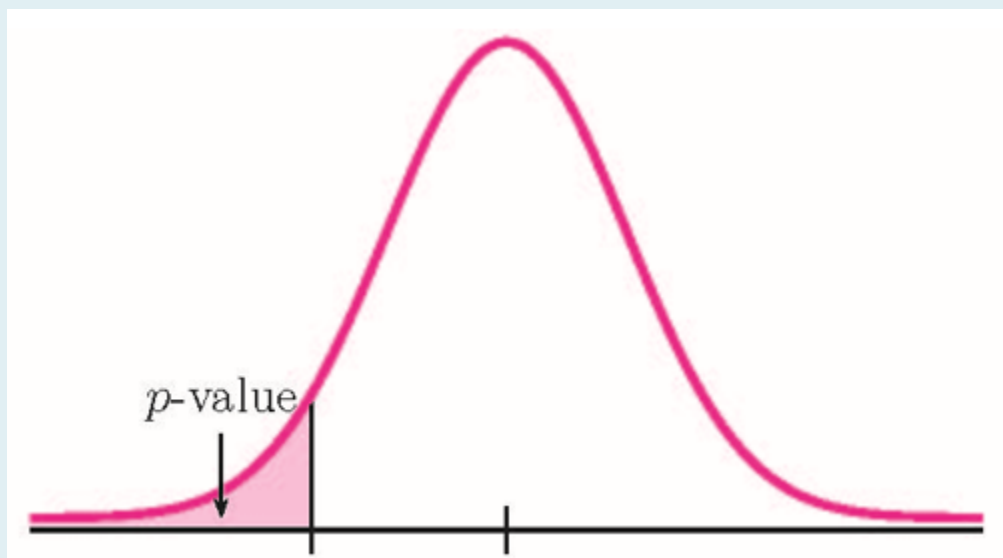
$$n_1 \times (1 - \hat{p}_1) = 100 \times (1 - 0.06) = 94 \geq 5$$

$$n_2 \times \hat{p}_2 = 150 \times 0.1066\dots = 16 \geq 5$$

$$n_2 \times (1 - \hat{p}_2) = 150 \times (1 - 0.1066\dots) = 134 \geq 5$$

Because $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1 - \hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$, and $n_2 \times (1 - \hat{p}_2) \geq 5$, the normal distribution applies, and so we use a normal distribution to calculate the p – value.

Because the alternative hypothesis is a $<$, the p – value is the area in the left tail of the distribution.



The pooled sample proportion is:

$$\begin{aligned}\bar{p} &= \frac{x_1 + x_2}{n_1 + n_2} \\ &= \frac{6 + 16}{100 + 150} \\ &= \frac{22}{250} \\ &= 0.088\end{aligned}$$

Function	norm.dist
Field 1	0.06-0.1066...
Field 2	0
Field 3	sqrt(0.088*(1-0.088)*(1/100+1/150))
Field 4	true
Answer	0.1010

So the p – value = 0.1010.

Conclusion:

Because p – value = 0.1010 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level, there is not enough evidence to suggest that the proportion of Valve As that fail under 4500 psi less than the proportion of Valve Bs that fail under 4500 psi.

NOTES

1. The null hypothesis $p_1 - p_2 = 0$ is the claim that the proportion of valves that fail under 4500 psi is the same for both valves ($p_1 = p_2$). That is, the two populations have the same proportion.
2. The alternative hypothesis $p_1 - p_2 < 0$ is the claim that the proportion of Valve A's that fail under 4500 psi less than the proportion of Valve B's that fail under 4500 psi ($p_1 < p_2$).
3. Make sure to keep all of the decimal places throughout the calculation to avoid any round-off error in the p – value. Perform the calculations of the sample proportions and the pooled sample proportion \bar{p} in Excel and then click on the corresponding cells when completing the

fields in the **norm.dist** function.

4. The p — value of 0.1010 is a large probability compared to the significance level, and so is likely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the proportion of Valve A's that fail under 4500 psi equals the proportion of Valve B's that fail under 4500 psi. For the company, this means that they could not claim that the proportion of their valves that fail under 4500 psi is the smallest of all the other valves on the market.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=227#oembed-1>

Video: “Excel 2013 Statistical Analysis #71: Inference About Difference Between 2 Pop. Proportions Z Method” by excelisfun [28:04] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A forestry researcher wants to compare the proportion of conifers in two different parts of the country. In a random sample of 100 forests in the western part of the country, 56 were coniferous or contained conifers. In a random sample of 80 forests in the east part of the country, 40 were coniferous or contained conifers. At the 5% significance level, is the proportion of conifers in the west part of the country greater than the proportion of conifers in the east part of the country?

Click to see Answer

Let the western part of the country be population 1 and the eastern part of the country be population 2.

- Hypotheses: $H_0 : p_1 - p_2 = 0$
 $H_a : p_1 - p_2 > 0$
- p - value = 0.2113
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the proportion of conifers in the west part of the country is not greater than the proportion of conifers in the east part of the country.

2. Two types of phone operating system are being tested to determine if there is a difference in the proportions of system failures (crashes). In a sample of 150 phones with OS_1 , 16 had system failures within the first eight hours of operation. In a sample of 150 phones with OS_2 , 8 had system failures within the first eight hours of operation. At the 5% significance level, is there a difference in the proportion of system failures for the two operating systems?

Click to see Answer

Let OS_1 be population 1 and OS_2 be population 2.

- Hypotheses: $H_0 : p_1 - p_2 = 0$
 $H_a : p_1 - p_2 \neq 0$
- p - value = 0.0887
- Conclusion: At the 5% significance level, there is enough evidence to conclude that there is no difference in the proportion of system failures for the two operating systems.

3. A local school district wants to compare the proportion of local high school seniors who used drugs or alcohol within the past month to the proportion of high school seniors nationwide. In a sample of 100 national high school seniors, 55 reported using drugs or alcohol within the past month. In a sample of 100 local high school seniors, 67 reported using drugs or alcohol within the past month. At the 5% significance level, is the proportion of nationwide high school seniors who used drugs or alcohol in the past month lower than the proportion of local high school seniors?

Click to see Answer

Let the nationwide seniors be population 1, and the local seniors be population 2.

- Hypotheses: $H_0 : p_1 - p_2 = 0$
 $H_a : p_1 - p_2 < 0$
- p - value = 0.0410
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the proportion of nationwide high school seniors who used drugs or alcohol in the past month is lower than the proportion of local high school seniors.

4. Neuroinvasive West Nile virus is a severe form of the West Nile virus that affects a person's nervous system. It is spread by the Culex species of mosquito. Last year, there were 629 reported cases of neuroinvasive West Nile virus out of a total of 1,021 reported West Nile cases. This year, there were 486 neuroinvasive reported cases out of a total of 712 reported West Nile cases. At the 1% significance level, is the proportion of neuroinvasive West Nile cases greater this year than last year?

Click to see Answer

Let last year be population 1 and this year be population 2.

- Hypotheses: $H_0 : p_1 - p_2 = 0$
 $H_a : p_1 - p_2 < 0$
- p - value = 0.0022
- Conclusion: At the 1% significance level, there is enough evidence to conclude that the proportion of neuroinvasive West Nile cases this year is greater than last year.

5. A researcher wants to study obesity in adults aged 18 years old and older. Adults are considered obese if their body mass index (BMI) is at least 30. In a sample of 248,775 women, 67,169 were considered obese. In a sample of 155,525 men, 42,769 were considered obese. At the 1% significance level, is the proportion of men who are obese greater than the proportion of women who are obese?

Click to see Answer

Let men be population 1 and women be population 2.

- Hypotheses: $H_0 : p_1 - p_2 = 0$
 $H_a : p_1 - p_2 > 0$
- p - value = 0.0003

- Conclusion: At the 1\% significance level, there is enough evidence to conclude that the proportion of men who are obese is greater than the proportion of women who are obese.

6. Two computer users were discussing tablet computers. One user claims that the proportion of people under 30 years old who use a tablet is the same as the proportion of people who are 30 years old or older. In a sample of 628 people under the age of 30 years, 603 said they use a tablet. In a sample of 2,309 people aged 30 years or older, 2,251 said they use a tablet. At the 1\% significance level, is the proportion of people under the age of 30 years who use a tablet different from the proportion of people aged 30 years or older who use a tablet?

Click to see Answer

Let under 30 years be population 1 and 30 years or older be population 2.

- Hypotheses: $H_0 : p_1 - p_2 = 0$
 $H_a : p_1 - p_2 \neq 0$
- p - value = 0.0489
- Conclusion: At the 1\% significance level, there is enough evidence to conclude that there is no difference in the proportion of people in the two age groups who own a tablet.

7. A group of friends debated whether more men use smartphones than women. They consulted a research study of smartphone use among adults. The results of the survey indicate that of the 973 men randomly sampled, 905 use smartphones. For women, 1,115 of the 1,304 who were randomly sampled use smartphones.
- Construct a 93\% confidence interval for the difference in the proportion of men and women who use smartphones.
 - Interpret the confidence interval in part (a).
 - Is it reasonable to claim that the proportion of men who use smartphones is higher than the proportion of women? Explain.

Click to see Answer

Let men be population 1 and women be population 2.

- Lower Limit = 0.0226 Upper Limit = 0.0662
- There is a 93\% probability that the difference in the proportions of men and women who use smartphones is between 2.26\% and 6.62\%.
- Yes. Because both limits are positive, it suggests that the difference in the proportions is positive. That is $p_1 - p_2 > 0$ or $p_1 > p_2$. So the proportion of men who use

smartphones is higher than the proportion of women.

8. Joan Nguyen recently claimed that the proportion of college-age males with at least one pierced ear is less than the proportion of college-age females. She conducted a survey in her classes. Out of 107 males, 20 had at least one pierced ear. Out of 92 females, 47 had at least one pierced ear.
- Construct a 98% confidence interval for the difference in the proportion of college-age males and females with at least one pierced ear.
 - Interpret the confidence interval in part (a).
 - Is it reasonable to claim that the proportion of college-age males with at least one pierced ear is less than the proportion of college-age females? Explain.

Click to see Answer

Let males be population 1 and females be population 2.

- Lower Limit = -0.4736 Upper Limit = -0.1743
- There is a 98% probability that the difference in the proportions of males and females with at least one pierced ear is between -47.36% and -17.43% .
- No. Because both limits are negative, it suggests that the difference in the proportions is negative. That is $p_1 - p_2 < 0$ or $p_1 < p_2$. So, the proportion of males with at least one pierced ear is less than the proportion of females with at least one pierced ear.

9. A business professor wants to compare the proportion of business students who work at the same time they attend school with the proportion of non-business students who work at the same time they attend school. In a sample of 160 business students, 110 said they work while attending school. In a sample of 160 non-business students, 115 said they work while attending school.
- Construct a 95% confidence interval for the difference in the proportion of business students and non-business students who work while attending school.
 - Interpret the confidence interval in part (a).
 - Can the professor claim that the proportions of business and non-business students who work while attending school are different? Explain.

Click to see Answer

Let business students be population 1 and non-business students be population 2.

- Lower Limit = -0.1313 Upper Limit = 0.688
- There is a 95% probability that the difference in the proportions of business students

and non-business students who work while attending school is between -13.13\% and 6.88\%.

- c. No. Because 0 is inside the confidence interval, it suggests that the difference in the proportions is 0. That is $p_1 - p_2 = 0$ or $p_1 = p_2$. So the proportion of business students who work while attending school is the same as the proportion of non-business students.

“9.5 Statistical Inference for Two Population Proportions” and “9.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

PART X

STATISTICAL INFERENCES USING THE CHI-SQUARE DISTRIBUTION

Previously, we looked at confidence intervals and hypothesis tests involving population means and population proportions, both single-population and two-populations scenarios. But what about other population parameters, such as the population variance? Like the other population parameters, we want to construct a confidence interval or conduct a hypothesis test for a single population variance. Unlike the population mean or the population variance, which use the normal or t -distributions, to study the population variance, we need to use a new distribution called the χ^2 -distribution.

Other situations use the χ^2 -distribution. For example, have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about whether a coffee machine was dispensing approximately the same amount of coffee each time? These situations involve testing how well an observed distribution of data fits the expected distribution. These types of scenarios require a hypothesis test using the χ^2 -distribution.

CHAPTER OUTLINE

10.1 The χ^2 -Distribution

10.2 Statistical Inference for a Single Population Variance

10.3 The Goodness-of-Fit Test

10.4 The Test of Independence

“10.1 Introduction to Statistical Inferences Using the Chi-Square Distribution” from Introduction

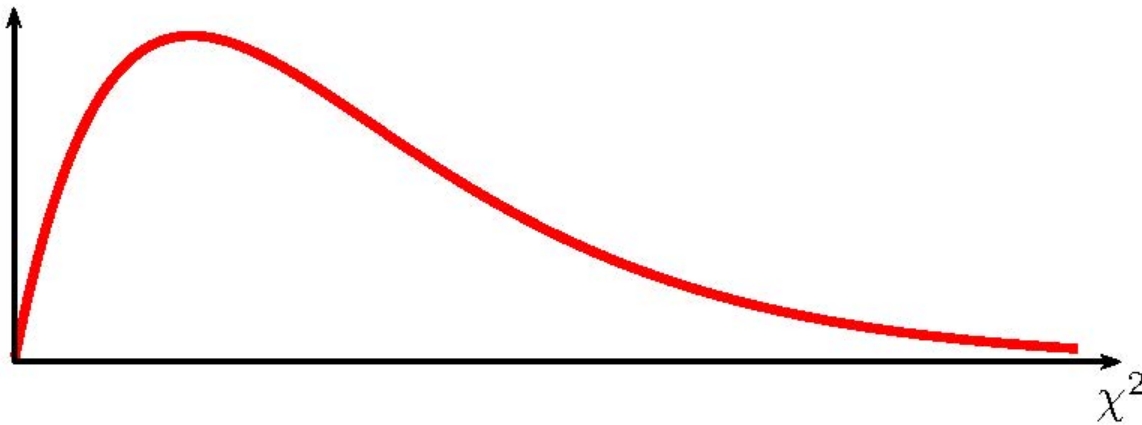
to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

10.1 THE CHI SQUARE DISTRIBUTION

LEARNING OBJECTIVES

- Find the area under a χ^2 -distribution.
- Find the χ^2 -score for a given area under the curve of a χ^2 -distribution.

The χ^2 -distribution is a continuous probability distribution. The graph of a χ^2 -distribution is shown below.



Properties of the χ^2 -distribution:

- The graph of a χ^2 -distribution is positively skewed and asymmetrical with a minimum value of 0 and no maximum value.
- A χ^2 -distribution is determined by its degrees of freedom, df . The value of the degrees of

freedom depends on how the χ^2 -distribution is used. There is a different χ^2 -distribution for every value of df . As the degrees of freedom increase, the χ^2 -distribution approaches a normal distribution.

- The total area under the graph of a χ^2 -distribution is 1.
- The mean of a χ^2 -distribution is its degrees of freedom: $\mu = df$.
- The variance of a χ^2 -distribution is twice its degrees of freedom: $\sigma^2 = 2 \times df$.
- The mode of a χ^2 -distribution is $df - 2$. The peak of the graph occurs at the mode.
- Probabilities associated with a χ^2 -distribution are given by the area under the curve of the χ^2 -distribution.

USING EXCEL TO CALCULATE THE AREA UNDER A χ^2 -DISTRIBUTION

To find the area in the left tail:

- To find the area under a χ^2 -distribution to the left of a given χ^2 -score, use the **chisq.dist**(χ^2 , **degrees of freedom**, **logic operator**) function.
 - For χ^2 , enter the χ^2 -score.
 - For **degrees of freedom**, enter the value of the degrees of freedom for the χ^2 -distribution.
 - For **logic operator**, enter **true**.
- The output from the **chisq.dist** function is the area to the left of the entered χ^2 -score.
- Visit the Microsoft page for more information about the **chisq.dist** function.

To find the area in the right tail:

- To find the area under a χ^2 -distribution to the right of a given χ^2 -score, use the **chisq.dist.rt**(χ^2 , **degrees of freedom**) function.
 - For χ^2 , enter the χ^2 -score.
 - For **degrees of freedom**, enter the value of the degrees of freedom for the χ^2 -distribution.

- The output from the **chisq.dist.rt** function is the area to the right of the entered χ^2 -score.
- Visit the Microsoft page for more information about the **chisq.dist.rt** function.

EXAMPLE

Consider a χ^2 -distribution with 12 degrees of freedom.

1. Find the area under the χ^2 -distribution to the left of $\chi^2 = 3.71$.
2. Find the area under the χ^2 -distribution to the right of $\chi^2 = 6.29$.

Solution

1.	Function	chisq.dist
	Field 1	3.71
	Field 2	12
	Field 3	true
	Answer	0.0119

2.	Function	chisq.dist.rt
	Field 1	6.72
	Field 2	12
	Answer	0.8755

USING EXCEL TO CALCULATE χ^2 -SCORES

To find the χ^2 -score for a given left-tail area:

- To find the χ^2 -score for a given area under the χ^2 -distribution to the left of the χ^2 -score, use the **chisq.inv(area to the left, degrees of freedom)** function.
 - For **area to the left**, enter the area to the left of required χ^2 -score.
 - For **degrees of freedom**, enter the value of the degrees of freedom for the χ^2 -distribution.
- The output from the **chisq.inv** function is the value of the χ^2 -score so that the area to the left of the χ^2 -score is the entered area.
- Visit the Microsoft page for more information about the **chisq.inv** function.

To find the χ^2 -score for a given right-tail area:

- To find the χ^2 -score for a given area under the χ^2 -distribution to the right of the χ^2 -score, use the **chisq.inv.rt(area to the right, degrees of freedom)** function.
 - For **area to the right**, enter the area to the right of required χ^2 -score.
 - For **degrees of freedom**, enter the value of the degrees of freedom for the χ^2 -distribution.
- The output from the **chisq.inv.rt** function is the value of the χ^2 -score so that the area to the right of the χ^2 -score is the entered area.
- Visit the Microsoft page for more information about the **chisq.inv.rt** function.

EXAMPLE

Consider a χ^2 -distribution with 37 degrees of freedom.

1. Find the χ^2 -score so that the area under the χ^2 -distribution to the left of χ^2 is 0.25.
2. Find the χ^2 -score so that the area under the χ^2 -distribution to the right of χ^2 is 0.148.

Solution

1.	Function	chisq.inv
	Field 1	0.25
	Field 2	37
	Answer	30.89

2.	Function	chisq.dist.rt
	Field 1	0.148
	Field 2	37
	Answer	45.97

TRY IT

Consider a χ^2 -distribution with 28 degrees of freedom.

1. Find the area under the χ^2 -distribution to the right of $\chi^2 = 21.7$.
2. Find the χ^2 -score so that area under the χ^2 -distribution to the left of χ^2 is 0.3.
3. Find the χ^2 -score so that area under the χ^2 -distribution to the right of χ^2 is 0.42.

4. Find the area under the χ^2 -distribution to the left of $\chi^2 = 17.3$.

Click to see Solution

1.

Function	chisq.dist.rt
Field 1	21.7
Field 2	28
Answer	0.795

2.

Function	chisq.inv
Field 1	0.3
Field 2	28
Answer	23.65

3.

Function	chisq.inv.rt
Field 1	0.42
Field 2	28
Answer	28.85

4.

Function	chisq.dist
Field 1	17.3
Field 2	28
Field 3	true
Answer	0.0576

Exercises

1. If the number of degrees of freedom for a χ^2 -distribution is 25, what is the population mean and standard deviation?

Click to see Answer

mean = 25, standard deviation = 7.07

2. Where is mode located on a χ^2 -distribution curve?

Click to see Answer

At the peak of the curve.

3. The variance of a χ^2 -distribution is 36. What is the mode?

Click to see Answer

16

4. Consider a χ^2 -distribution with 17 degrees of freedom.

- Find the area under the χ^2 -distribution to the left of $\chi^2 = 15.3$.
- Find the area under the χ^2 -distribution to the right of $\chi^2 = 22.8$.
- Find the χ^2 -score so that area under the χ^2 -distribution to the left of χ^2 is 0.291.
- Find the χ^2 -score so that area under the χ^2 -distribution to the right of χ^2 is 0.3156.

Click to see Answer

- 0.426
- 0.1559
- 13.4
- 19.23

5. Consider a χ^2 -distribution with 12 degrees of freedom.

- What is the mean of the χ^2 -distribution?
- What is the mode of the χ^2 -distribution?
- What is the variance of the χ^2 -distribution?
- Find the area under the χ^2 -distribution to the left of $\chi^2 = 82$.
- Find the area under the χ^2 -distribution to the right of $\chi^2 = 14.9$.
- Find the χ^2 -score so that area under the χ^2 -distribution to the left of χ^2 is 0.1183.
- Find the χ^2 -score so that area under the χ^2 -distribution to the right of χ^2 is 0.6977.

Click to see Answer

- 12
- 10
- 24

- d. 0.2307
- e. 0.2470
- f. 6.62
- g. 9.06

“10.2 The Chi Square Distribution” and “10.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

10.2 STATISTICAL INFERENCE FOR A SINGLE POPULATION VARIANCE

LEARNING OBJECTIVES

- Calculate and interpret a confidence interval for a population variance.
- Conduct and interpret a hypothesis test on a single population variance.

The mean of a population is important, but in many cases, the variance of the population is just as important. In most production processes, quality is measured by how closely the process matches the target (i.e. the mean) and by the variability (i.e. the variance) of the process. For example, if a process is to fill bags of coffee beans, we are interested in both the mean weight of the bag and how much variation there is in the weight of the bags. The quality is considered poor if the mean weight of the bags is accurate, but the variance of the weight of the bags is too high—a variance that is too large means some bags would be too full, and some bags would be almost empty.

As with other population parameters, we can construct a confidence interval to capture the population variance and conduct a hypothesis test on the population variance. In order to construct a confidence interval or conduct a hypothesis test on a population variance σ^2 , we need to use the distribution of $\frac{(n-1) \times s^2}{\sigma^2}$. Suppose we have a normal population with population variance σ^2 , and a sample of size n is taken from the population. The sampling distribution of $\frac{(n-1) \times s^2}{\sigma^2}$ follows a χ^2 -distribution with $n - 1$ degrees of freedom.

Constructing a Confidence Interval for a Population Variance

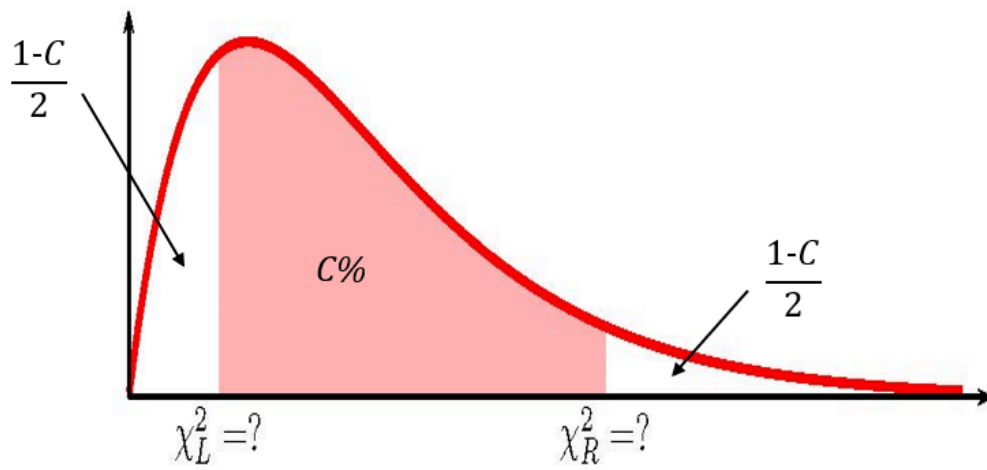
To construct the confidence interval, take a random sample of size n from a normally distributed

population. Calculate the sample variance s^2 . The limits for the confidence interval with confidence level C for an unknown population variance σ^2 are

$$\text{Lower Limit} = \frac{(n - 1) \times s^2}{\chi_R^2}$$

$$\text{Upper Limit} = \frac{(n - 1) \times s^2}{\chi_L^2}$$

where χ_L^2 is the χ^2 -score so that the area in the left tail of the χ^2 -distribution is $\frac{1 - C}{2}$, χ_R^2 is the χ^2 -score so that the area in the right tail of the χ^2 -distribution is $\frac{1 - C}{2}$ and the χ^2 -distribution has $n - 1$ degrees of freedom.



NOTES

1. Like the other confidence intervals we have seen, the χ^2 -scores are the values that trap $C\%$ of the observations in the middle of the distribution so that the area of each tail is $\frac{1 - C}{2}$.

2. Because the χ^2 -distribution is not symmetrical, the confidence interval for a population variance requires that we calculate **two** different χ^2 -scores: one for the left tail and one for the right tail. In Excel, we need to use both the **chisq.inv** function (for the left tail) and the **chisq.inv.rt** function (for the right tail) to find the two different χ^2 -scores.
3. The χ^2 -score for the left tail is part of the formula for the upper limit, and the χ^2 -score for the right tail is part of the formula for the lower limit. **This is not a mistake.** It follows from the formula used to determine the limits for the confidence interval.

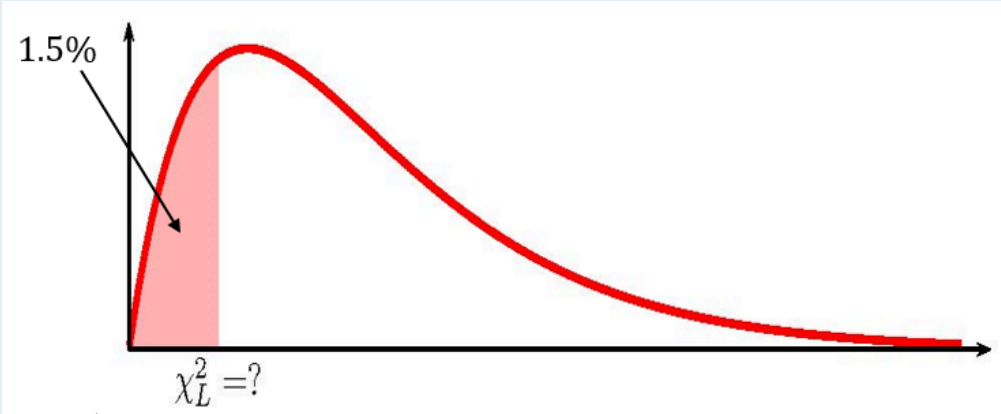
EXAMPLE

A local telecom company conducts broadband speed tests to measure how much data per second passes between a customer's computer and the internet compared to what the customer pays for as part of their plan. The company needs to estimate the variance in the broadband speed. A sample of 15 ISPs is taken, and the amount of data per second is recorded. The variance in the sample is 174.

1. Construct a 97% confidence interval for the variance in the amount of data per second that passes between a customer's computer and the internet.
2. Interpret the confidence interval found in part 1.

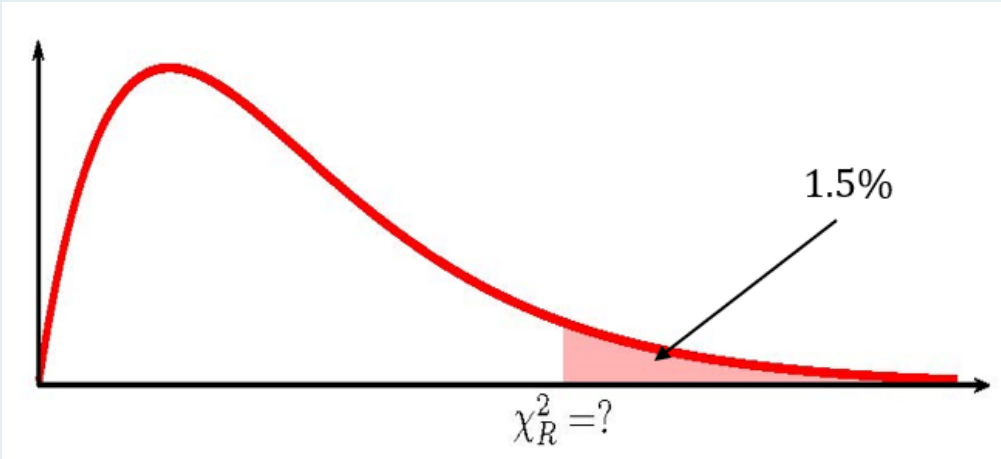
Solution

1. To find the confidence interval, we need to find the χ_L^2 -score for the 97% confidence interval. This means that we need to find the χ_L^2 -score so that the area in the left tail is $\frac{1 - 0.97}{2} = 0.015$. The degrees of freedom for the χ^2 -distribution is $n - 1 = 15 - 1 = 14$.



Function	chisq.inv
Field 1	0.015
Field 2	14
Answer	5.0572...

We also need to find the χ_R^2 -score for the 97\% confidence interval. This means that we need to find the χ_R^2 -score so that the area in the right tail is $\frac{1 - 0.97}{2} = 0.015$. The degrees of freedom for the χ^2 -distribution is $n - 1 = 15 - 1 = 14$.



Function	chisq.inv.rt
Field 1	0.015
Field 2	14
Answer	27.826...

So $\chi_L^2 = 5.0572 \dots$ and $\chi_R^2 = 27.826 \dots$. From the sample data supplied in the question $s^2 = 174$ and $n = 15$. The 97% confidence interval is

$$\begin{aligned}\text{Lower Limit} &= \frac{(n-1) \times s^2}{\chi_R^2} \\ &= \frac{(15-1) \times 174}{27.826 \dots} \\ &= 87.54\end{aligned}$$

$$\begin{aligned}\text{Upper Limit} &= \frac{(n-1) \times s^2}{\chi_L^2} \\ &= \frac{(15-1) \times 174}{5.0572 \dots} \\ &= 481.69\end{aligned}$$

1. We are 97% confident that the variance in the amount of data per second that passes between a customer's computer and the internet is between 87.54 and 481.69.

NOTES

1. When calculating the limits for the confidence interval, keep all of the decimals in the χ^2 -scores and other values throughout the calculation. This will ensure that there is no round-off error in the answer. Use Excel to do the calculations of the limits, clicking on the cells containing the χ^2 -scores and any other values.
2. When writing down the interpretation of the confidence interval, make sure to include the confidence level and the actual population variance captured by the confidence interval (i.e. be specific to the context of the question).

TRY IT

A pharmaceutical company manufactures a particular drug. To ensure that patients receive the proper dose, the variance in the weight of the drug is critical. The company's quality control department routinely assesses the drug as it is produced to ensure it meets production guidelines. From the latest production run, a sample of 40 units of the drug is taken, and the weight, in grams, of each unit is recorded. The variance of the weights is 0.27

1. Construct a 95% confidence interval for the variance in the weight of the drug.
2. Interpret the confidence interval found in part 1.
3. If the variance in the weight of the drug exceeds 0.5, the drug is rejected. Based on the sample, should the quality control department reject the latest production run of the drug? Explain.

Click to see Solution

1.

Function	chisq.inv
Field 1	0.025
Field 2	39
Answer	23.654...

Function	chisq.inv.rt
Field 1	0.025
Field 2	39
Answer	58.120...

$$\begin{aligned}
 \text{Lower Limit} &= \frac{(n-1) \times s^2}{\chi_R^2} \\
 &= \frac{(40-1) \times 0.27}{58.120\dots} \\
 &= 0.181
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \frac{(n-1) \times s^2}{\chi_L^2} \\
 &= \frac{(40-1) \times 0.27}{23.654\dots} \\
 &= 0.445
 \end{aligned}$$

2. We are 95\% confident that the variance in the weight of the drug is between 0.181 and 0.445.
3. The quality control department should not reject the latest production run of the drug because 0.5 is outside the confidence interval. The variance of the weight of the drug does not exceed 0.5.

Conducting a Hypothesis Test for a Population Variance

Follow these steps to perform a hypothesis test for a population variance:

1. Write down the null and alternative hypotheses in terms of the population variance σ^2 .
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information for the test and identify the significance level α .
4. Use the χ^2 -distribution to find the p – value (the area in the corresponding tail) for the test. The χ^2 -score and degrees of freedom are

$$\chi^2 = \frac{(n-1) \times s^2}{\sigma^2} \quad df = n - 1$$

5. Compare the p – value to the significance level and state the outcome of the test.

- If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
- If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is insufficient evidence to conclude that the alternative hypothesis H_a may be correct.

6. Write down a concluding sentence specific to the context of the question.

EXAMPLE

A statistics instructor at a local college claims that the variance for the final exam scores was 25. After speaking with his classmates, one of the class's best students thinks that the variance for the final exam scores is higher than the instructor claims. The student challenges the instructor to prove her claim. The instructor takes a sample 30 final exams and finds the variance of the scores is 28. At the 5\% significance level, test if the variance of the final exam scores is higher than the instructor claims.

Solution

Hypotheses:

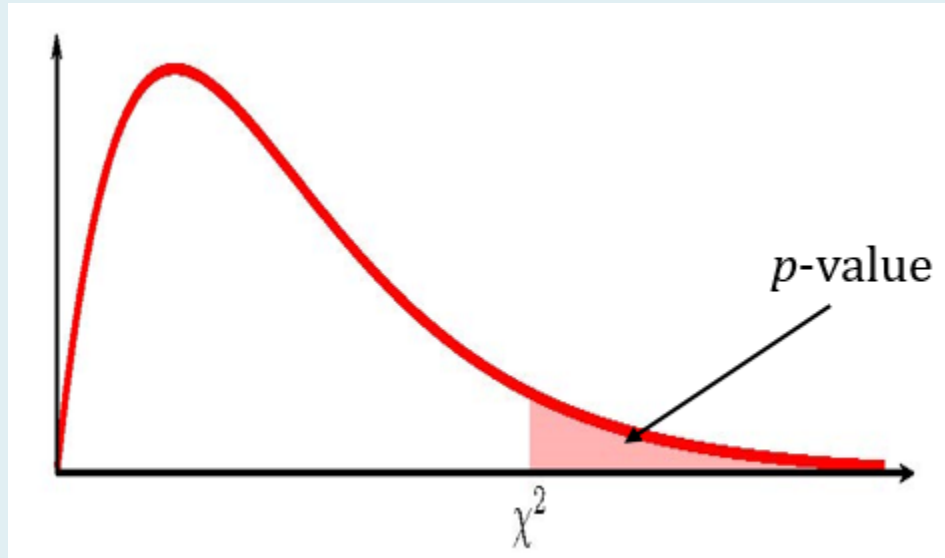
$$H_0 : \sigma^2 = 25$$

$$H_a : \sigma^2 > 25$$

p – value:

From the question, we have $n = 30$, $s^2 = 28$, and $\alpha = 0.05$.

Because the alternative hypothesis is a $>$, the p -value is the area in the right tail of the χ^2 -distribution.



To use the **chisq.dist.rt** function, we need to calculate out the χ^2 -score and the degrees of freedom:

$$\begin{aligned}\chi^2 &= \frac{(n - 1) \times s^2}{\sigma^2} \\ &= \frac{(30 - 1) \times 28}{25} \\ &= 32.48\end{aligned}$$

$$\begin{aligned}df &= n - 1 \\ &= 30 - 1 \\ &= 29\end{aligned}$$

Function	chisq.dist.rt
Field 1	32.48
Field 2	29
Answer	0.2992

So the p – value = 0.2992.

Conclusion:

Because p – value = 0.2992 > 0.05 = α , we do not reject the null hypothesis. At the 5\%

significance level, there is not enough evidence to suggest that the variance of the final exam scores is higher than 25.

NOTES

1. The null hypothesis $\sigma^2 = 25$ is the claim that the variance on the final exam is 25.
2. The alternative hypothesis $\sigma^2 > 25$ is the claim that the variance on the final exam is greater than 25.
3. The p – value is the area in the right tail of the χ^2 -distribution, to the right of $\chi^2 = 32.84$. In the calculation of the p – value:
 - The function is **chisq.dist.rt** because we are finding the area in the right tail of a χ^2 -distribution.
 - Field 1 is the value of χ^2 .
 - Field 2 is the degrees of freedom.
4. The p – value of 0.2992 is a large probability compared to the significance level and so is likely to happen, assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the variance of the scores on the final exam is most likely 25.

EXAMPLE

With individual lines at its various windows, a post office finds that the standard deviation for normally distributed waiting times for customers is 7.2 minutes. The post office experiments with a single, main waiting line and finds that for a random sample of 25 customers, the waiting times for customers have a standard deviation of 4.5 minutes. At the 5% significance level, determine if the single line changed the variation among the wait times for customers.

Solution**Hypotheses:**

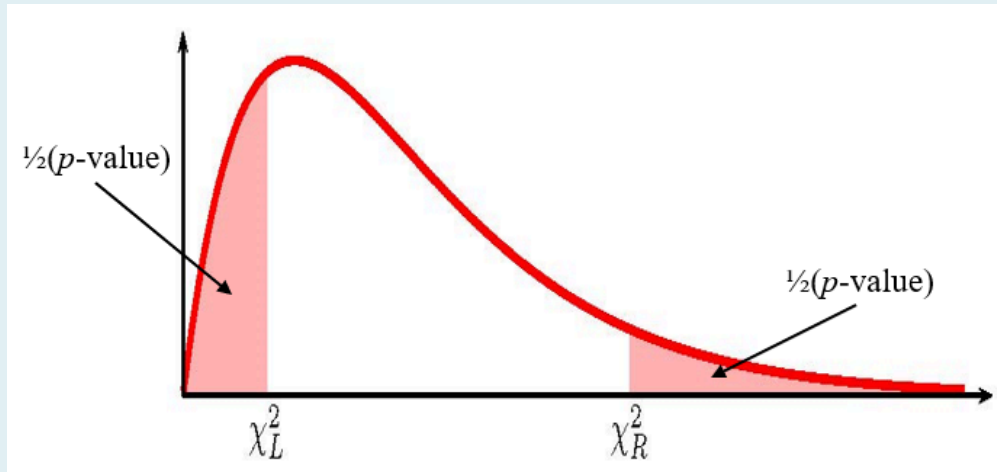
$$H_0 : \sigma^2 = 51.84$$

$$H_a : \sigma^2 \neq 51.84$$

p – value:

From the question, we have $n = 25$, $s^2 = 20.25$, and $\alpha = 0.05$.

Because the alternative hypothesis is a \neq , the p – value is the sum of the areas in the tails of the χ^2 -distribution.



We need to calculate out the χ^2 -score and the degrees of freedom:

$$\begin{aligned}\chi^2 &= \frac{(n - 1) \times s^2}{\sigma^2} \\ &= \frac{(25 - 1) \times 20.25}{51.84} \\ &= 9.375\end{aligned}$$

$$\begin{aligned}df &= n - 1 \\ &= 25 - 1 \\ &= 24\end{aligned}$$

Because this is a two-tailed test, we need to know which tail (left or right) the χ^2 -score belongs to so that we can use the correct Excel function. If $\chi^2 > df - 2$, the χ^2 -score corresponds to the right tail. If the $\chi^2 < df - 2$, the χ^2 -score corresponds to the left tail. In this case,

$\chi^2 = 9.375 < 22 = df - 2$, so the χ^2 -score corresponds to the left tail. We need to use **chisq.dist** to find the area in the left tail.

Function	chisq.dist
Field 1	9.375
Field 2	24
Answer	0.0033

So the area in the left tail is 0.0033, which means that $\frac{1}{2}(p - \text{value}) = 0.0033$. This is also the area in the right tail, so

$$p - \text{value} = 0.0033 + 0.0033 = 0.0066$$

Conclusion:

Because $p - \text{value} = 0.0066 < 0.05 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level, there is enough evidence to suggest that the variation among the wait times for customers has changed.

NOTES

1. The null hypothesis $\sigma^2 = 51.84$ is the claim that the variance in the wait times is 51.84. Note that we were given the standard deviation ($\sigma = 7.2$) in the question. But this is a test on variance, so we must write the hypotheses in terms of the variance
 $\sigma^2 = 7.2^2 = 51.84$.
2. The alternative hypothesis $\sigma^2 \neq 51.84$ is the claim that the variance in the wait times has changed from 51.84.
3. In a two-tailed hypothesis test for population variance, we will only have sample information relating to **one** of the two tails. We must determine which of the tails the sample information belongs to, and then calculate out the area in that tail. The area in each tail represents exactly half of the $p - \text{value}$, so the $p - \text{value}$ is the sum of the areas in the two tails.
 - If $\chi^2 < df - 2$, the sample information belongs to the **left tail**.
 - We use **chisq.dist** to find the area in the left tail. The area in the right tail equals the area in the left tail, so we can find the $p - \text{value}$ by adding the output

from this function to itself.

- If $\chi^2 > df - 2$, the sample information belongs to the **right tail**.
 - We use **chisq.dist.rt** to find the area in the right tail. The area in the left tail equals the area in the right tail, so we can find the **p — value** by adding the output from this function to itself.
- 4. The **p — value** of **0.0066** is a small probability compared to the significance level and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the variance in the wait times has most likely changed.

TRY IT

A scuba instructor wants to record the collective depths each of his students dives during their checkout. He is interested in how the depths vary, even though everyone should have been at the same depth. He believes the standard deviation of the depths is **1.2** meters. But his assistant thinks the standard deviation is less than **1.2** meters. The instructor wants to test this claim. The scuba instructor uses his most recent class of **20** students as a sample and finds that the standard deviation of the depths is **0.85** meters. At the **1\%** significance level, test if the variability in the depths of the student scuba divers is less than claimed.

Click to see Solution

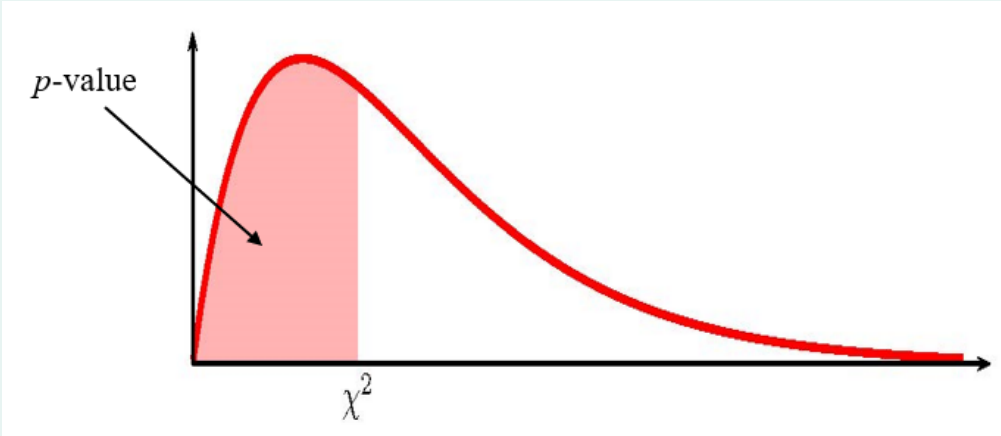
Hypotheses:

$$H_0 : \sigma^2 = 1.44$$
$$H_a : \sigma^2 < 1.44$$

p – value:

From the question, we have $n = 20$, $s^2 = 0.7225$, and $\alpha = 0.01$.

Because the alternative hypothesis is a $<$, the *p* – value is the area in the left tail of the χ^2 -distribution.



To use the **chisq.dist** function, we need to calculate out the χ^2 -score and the degrees of freedom:

$$\begin{aligned}\chi^2 &= \frac{(n - 1) \times s^2}{\sigma^2} \\ &= \frac{(20 - 1) \times 0.7225}{1.44} \\ &= 9.5329 \dots\end{aligned}$$

$$\begin{aligned}df &= n - 1 \\ &= 20 - 1 \\ &= 19\end{aligned}$$

Function	chisq.dist
Field 1	9.5329...
Field 2	19
Field 3	true
Answer	0.0365

So the p – value = 0.0365.

Conclusion:

Because p – value = 0.0365 $>$ 0.01 = α , we do not reject the null hypothesis. At the 1\% significance level, there is not enough evidence to suggest that the variation in the depths of the students is less than claimed.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=244#oembed-1>

Video: “Hypothesis Tests for One Population Variance” by jbststatistics [8:52] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. An archer claims that the variance for his hits is 36 (data is measured in distance from the center of the target). An observer claims the archer’s standard deviation for his hits is less, and the observer wants to test her claim at the 5\% significance level. The observer takes a sample of 30 of the archer’s hits and finds a variance of 29.16.

Click to see Answer

- Hypotheses: $H_0 : \sigma^2 = 36$
 $H_a : \sigma^2 < 36$
- p – value = 0.2463
- Conclusion: At the 5\% significance level, there is not enough evidence to conclude that the archer’s variance is less than 36.

2. In the past, the variance of heights for students in a school is 0.66. A researcher believes the variation of the heights of students at the school is greater than 0.66. A random sample of 50 students at the school is taken, and the standard deviation of heights in the sample is 0.96. At the 5% significance level, determine if the variance in the heights for students in the school is greater than 0.66.

Click to see Answer

- Hypotheses: $H_0 : \sigma^2 = 0.66$
 $H_a : \sigma^2 > 0.66$
- p – value = 0.0347
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the variation in student heights at the school is greater than 0.66.

3. The average waiting time to see a doctor at a medical clinic varies. One doctor at the clinic wants to estimate the variance in the waiting times. In a random sample of 30 patients at the medical clinic, the standard deviation of waiting times of 4.1 minutes.
- a. Construct a 96% confidence interval for the variation in the wait times at the doctor's office.
 - b. Interpret the confidence interval found in part (a).
 - c. The doctor investigating the variance of wait times believes that the variance in the wait times is greater than 12. Is the doctor's claim reasonable? Explain.

Click to see Answer

- a. Lower Limit = 10.44, Upper Limit = 31.30
 - b. There is a 96% probability that the variance in the wait times is between 10.44 and 31.30.
 - c. Yes, because 12 is inside the confidence interval.
4. Suppose an airline claims that its flights are consistently on time with an average delay of at most 15 minutes. It claims that the average delay is so consistent that the variance is no more than 150. Doubting the consistency part of the claim, a disgruntled traveller calculates the delays for his next 25 flights. The average delay for those 25 flights is 22 minutes with a standard deviation of 15 minutes. At the 5% significance level, determine if the variance in the delay times is greater than 150.

Click to see Answer

- Hypotheses: $H_0 : \sigma^2 = 150$
 $H_a : \sigma^2 > 150$
- p – value = 0.0853
- Conclusion: At the 5\% significance level, there is not enough evidence to conclude that the variance in the delay times is greater than 150.

5. A plant manager at a cereal production company is concerned her equipment may need recalibrating. It seems that the actual weight of the 750 gram cereal boxes the equipment fills has been fluctuating. In order to determine if the machine needs to be recalibrated, 84 randomly selected boxes of cereal from the next day's production were weighed. The variance of the 84 boxes was 1.3.
- a. Construct a 98\% confidence interval for the variance in the weight of the cereal boxes.
 - b. Interpret the confidence interval found in part (a).
 - c. If the variance in the weight of the cereal boxes is supposed to be at most 2, does the machine need to be recalibrated?

Click to see Answer

- a. Lower Limit = 0.931, Upper Limit = 1.927
 - b. There is a 98\% probability that the variance in the weight of the cereal boxes is between 0.931 and 1.927.
 - c. No, because 2 is outside the confidence interval.
6. Consumers may be interested in whether the cost of a particular calculator varies from store to store. A major retailer claims that the variation in the price of the calculator is 225. But one consumer doubts this claim and believes that the variance in the price of the calculator is greater than 225. Based on a sample of 43 stores, the consumer found that the mean price of the calculator was \$84 and a standard deviation of \$18. At the 5\% significance level, test the consumer's claim that the variance in the price of the calculator is greater than 225.

Click to see Answer

- Hypotheses: $H_0 : \sigma^2 = 225$
 $H_a : \sigma^2 > 225$
- p – value = 0.0322
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that the variance in the price of the calculator is greater than 225.

7. Airline companies are interested in the consistency of the number of babies on each flight so that they have adequate safety equipment. They are also interested in the variation of the number of babies. Suppose that an airline executive believes the variance in the number of babies per flight is 9. The airline wants to test the executive's claim to see if the variance is different than claimed. The airline takes a sample of 18 flights and finds that the standard deviation for the number of babies per flight is 4.1. Conduct a hypothesis test of the airline executive's belief. Use a 1% significance level.

Click to see Answer

- Hypotheses: $H_0 : \sigma^2 = 9$
 $H_a : \sigma^2 \neq 9$
- p – value = 0.0323
- Conclusion: At the 1% significance level, there is not enough evidence to conclude that the variance in the number of babies per flight is different from 9.

8. According to an avid aquarist, the variance for the number of fish in a 20-gallon tank is 4. His friend, also an aquarist, does not believe this claim and that the variance is actually different from 4. She counts the number of fish in 15 other 20-gallon tanks and gets the following results:

11	11	11	9	11
10	10	10	7	10
9	10	12	9	11

At the 5% significance level, test the claim that the variance for the number of fish in a 20-gallon tank is different than 4.

Click to see Answer

- Hypotheses: $H_0 : \sigma^2 = 4$
 $H_a : \sigma^2 \neq 4$
- p – value = 0.0354
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the variance in the number of fish in a 20-gallon tank is different from 4.

9. The manager of “Frenchies” is concerned that patrons are not consistently receiving the same amount of French fries with each order. The chef claims that the variation for an order of fries is 2.25, but the manager thinks that it may be lower. He randomly weighs 49 orders of fries, which yields a variance of 1.49. At the 1% significance level, determine if the variation in the amount of fries per order is less than claimed.

Click to see Answer

- Hypotheses: $H_0 : \sigma^2 = 2.25$
 $H_a : \sigma^2 < 2.25$
- p – value = 0.0344
- Conclusion: At the 1% significance level, there is not enough evidence to conclude that the variance in the amount of fries per order is less than 2.25.

10. You want to buy a specific computer. A sales representative of the manufacturer claims that retail stores sell this computer at an average price of \$1,249 with a variance of 625. You find a website that has a price comparison for the same computer at a series of stores as follows:

\$1299	\$1299.99	\$1193.08	\$1279
1224.95	1229.99	1269.95	1249

Can you argue that the variation in the price of the computer is different than what is claimed by the manufacturer? Use the 5% significance level.

Click to see Answer

- Hypotheses: $H_0 : \sigma^2 = 625$
 $H_a : \sigma^2 \neq 625$
- p – value = 0.0459
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the variance in the price of the computer is greater than 625.

11. A company packages apples by weight. One of the weight grades is Class A apples. A batch of apples is selected to be included in a Class A apple package. The weights of the selected apples (in grams) is as follows:

158	167	149	169	164
139	154	150	157	171
152	161	141	166	172

- Construct a 95\% confidence interval for the variation in the weight of apples in the package.
- Interpret the confidence interval found in part (a).

Click to see Answer

- Lower Limit = 58.35, Upper Limit = 270.75
- There is a 95\% probability that the variance in the weight of the apples is between 58.35 and 270.75.

“10.3 Statistical Inference for a Single Population Variance” and “10.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

10.3 THE GOODNESS-OF-FIT TEST

LEARNING OBJECTIVES

- Conduct and interpret χ^2 -goodness-of-fit hypothesis tests.

Recall that a **categorical** (or qualitative) variable is a variable where the data can be grouped by specific categories. Examples of categorical variables include eye colour, blood type, or brand of car. A categorical variable is a random variable that takes on categories. Suppose we want to determine whether the data from a categorical variable “fit” a particular distribution or not. That is, for a categorical variable with a historical or assumed probability distribution, does a new sample from the population support the assumed probability distribution, or does the sample indicate that there has been a change in the probability distribution?

The χ^2 -goodness-of-fit test allows us to test if the sample data from a categorical variable fits the pattern of **expected probabilities** for the variable. In a χ^2 -goodness-of-fit test, we are analyzing the distribution of the frequencies for one categorical variable. This is a hypothesis test where the hypotheses state that the categorical variable does or does not follow an assumed probability distribution, and a χ^2 -distribution is used to determine the p – value for the test.

Conducting a χ^2 -Goodness-of-Fit Test

Suppose a categorical variable has k possible outcomes (categories) with probabilities p_1, p_2, \dots, p_k . Suppose n independent observations are taken from this categorical variable.

1. Write down the null and alternative hypotheses:

$$H_0 : p_1 = p_{1_0}, p_2 = p_{2_0}, \dots, p_k = p_{k_0}$$

$$H_a : \text{at least one } p_i \neq p_{i_0}$$

2. Collect the sample information for the test and identify the significance level α .
3. Use the χ^2 -distribution to find the p – value, which is the **area in the right tail** of the χ^2 -distribution. The χ^2 -score and degrees of freedom are

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$df = k - 1$$

observed = observed frequency from the sample data

expected = expected frequency from assumed distribution

k = number of categories

4. Compare the p – value to the significance level and state the outcome of the test.
 - If p – value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If p – value $> \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is no sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
5. Write down a concluding sentence specific to the context of the question.

NOTES

1. The null hypothesis is the claim that the categorical variable follows the assumed distribution. That is, the probability p_i of each possible outcome of the categorical variable equals a hypothesized probability p_{i_0} .

2. The alternative hypothesis is the claim that the categorical variable does not follow the assumed distribution. That is, for at least one possible outcome of the categorical variable, the probability p_i does not equal the claimed probability p_{i_0} .
3. In order to use the χ^2 -goodness-of-fit test, the expected frequency for each category must be at least 5.
4. The p — value for a χ^2 -goodness-of-fit test is always the area in the right tail of the χ^2 -distribution. So, we use **chisq.dist.rt** to find the p — value for a χ^2 -goodness-of-fit test.
5. To calculate the χ^2 -score:
 - For each of the possible outcomes of the categorical variable, calculate $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$:
 - i. Find the difference between the observed frequency (from the sample) and the expected frequency (from the null hypothesis). The expected frequency equals $n \times p_{i_0}$ where n is the sample size and p_{i_0} is the assumed probability for the i th outcome claimed in the null hypothesis.
 - ii. Square the difference in step (i).
 - iii. Divide the value found in step (ii) by the expected frequency.
 - Add up the values of $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ for each of the outcomes.
6. We expect that there will be a discrepancy between the observed frequency and the expected frequency. If this discrepancy is very large, the value of χ^2 will be very large and result in a small p — value.

EXAMPLE

Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the

actual student absenteeism rate follows faculty perception. The faculty believe that the distribution of the number of absences per term is as follows:

Number of Absences per Term	Expected Percent of Students
0–2	50\%
3–5	30\%
6–8	12\%
9–11	6\%
12+	2\%

At the end of the semester, a random survey of 300 students across all mathematics courses was taken, and the actual (observed) number of absences for the 300 students was recorded.

Number of Absences per Term	Observed Number of Students
0–2	120
3–5	100
6–8	55
9–11	15
12+	10

At the 5\% significance level, determine if the number of absences per term follows the distribution assumed by the faculty.

Solution

Let p_1 be the probability a student has 0-2 absences, p_2 be the probability a student has 3-5 absences, p_3 be the probability a student has 6-8 absences, p_4 be the probability a student has 9-11 absences, and p_5 be the probability a student has 12 or more absences.

Hypotheses:

$$\begin{array}{l} H_0: p_1=50\%, p_2=30\%, p_3=12\%, p_4=6\%, p_5=2\% \\ H_a: \text{at least one of the } p_i \text{ does not equal its stated probability} \end{array}$$

p – value:

From the question, we have $n = 300$ and $k = 5$. Now we need to calculate the χ^2 -score for the test.

The observed frequency for each category is the number of observations in the sample that fall into that category. This is the information provided in the sample above.

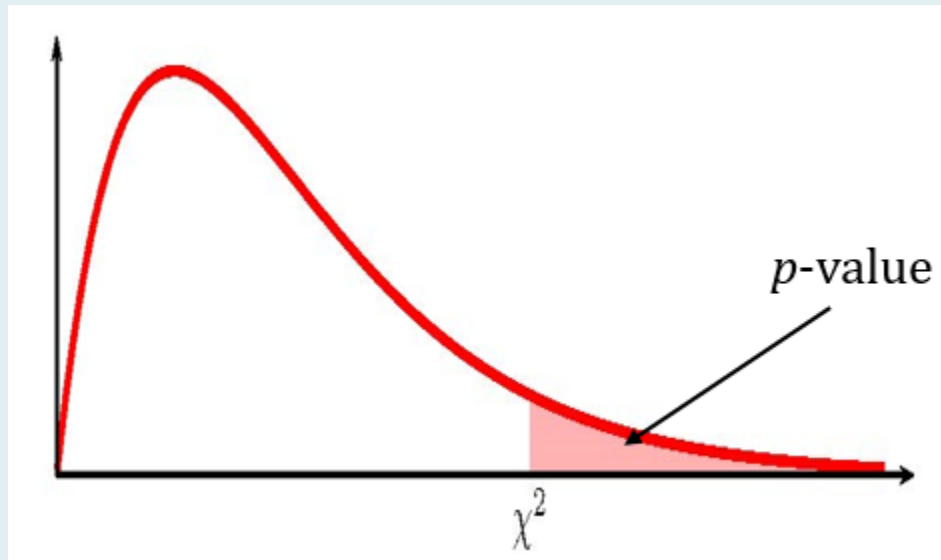
Next, we must calculate out the expected frequencies. The expected frequency is the number of observations we would expect to see in the sample, assuming the null hypothesis is true. To calculate the expected frequency for each category, we multiply the sample size n by the probability associated with that category claimed in the null hypothesis.

Number of Absences per Term	Observed Frequency	Expected Frequency
0-2	120	$0.5 \times 300 = 150$
3-5	100	$0.3 \times 300 = 90$
6-8	55	$0.12 \times 300 = 36$
9-11	15	$0.06 \times 300 = 18$
12+	10	$0.02 \times 300 = 6$

To calculate the χ^2 -score, we work out the quantity $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ for each category and then add up these quantities.

$$\begin{aligned}
 \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\
 &= \frac{(120 - 150)^2}{150} + \frac{(100 - 90)^2}{90} + \frac{(55 - 36)^2}{36} + \frac{(15 - 18)^2}{18} + \frac{(10 - 6)^2}{6} \\
 &= 20.305 \dots
 \end{aligned}$$

The degrees of freedom for the χ^2 -distribution is $df = k - 1 = 5 - 1 = 4$. The χ^2 -goodness-of-fit test is a right-tailed test, so we use the **chisq.dist.rt** function to find the p - value:



Function	chisq.dist.rt
Field 1	20.305....
Field 2	4
Answer	0.0004

So the p – value = 0.0004.

Conclusion:

Because p – value = 0.0004 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5\% significance level, there is enough evidence to suggest that the number of absences per term does not follow the distribution assumed by faculty.

NOTES

1. The null hypothesis is the claim that the percent of students that fall into each category is as stated. That is, 50\% students miss between 0 and 2 classes, 30\% of the students miss between 3 and 5 students, etc.
2. The alternative hypothesis is the claim that at least one of the percent of students that fall into each category is not as stated. The alternative hypothesis does not say that every p_i does not equal its stated probabilities, only that one of them does not equal its stated probability.

3. Keep all of the decimals throughout the calculation (i.e. in the calculation of the χ^2 -score) to avoid any round-off error in the calculation of the p — value. This ensures that we get the most accurate value for the p — value. Use Excel to calculate the expected frequencies and the χ^2 -score.
4. The p — value is the area in the right tail of the χ^2 -distribution, to the right of $\chi^2 = 20.305 \dots$. In the calculation of the p — value:
 - The function is **chisq.dist.rt** because we are finding the area in the right tail of a χ^2 -distribution.
 - Field 1 is the value of χ^2 .
 - Field 2 is the value of the degrees of freedom df .
5. The p — value of **0.0004** is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, student absenteeism does not fit faculty perception.

EXAMPLE

Employers want to know which days of the week employees have the highest number of absences in a five-day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers are asked on which day of the week they had the highest number of employee absences. The results are recorded in the table below. At the 5% significance level, test if the day of the week with the highest number of absences occurs with equal frequency during a five-day work week.

Day of the Week	Observed Frequency
Monday	15
Tuesday	11
Wednesday	10
Thursday	9
Friday	15

Solution

Let p_1 be the probability the highest number of absences occurs on Monday, p_2 be the probability the highest number of absences occurs on Tuesday, p_3 be the probability the highest number of absences occurs on Wednesday, p_4 be the probability the highest number of absences occurs on Thursday, and p_5 be the probability the highest number of absences occurs on Friday.

If the day of the week with the highest number of absences occurs with equal frequency, then the probability that any day has the highest number of absences is the same as any other day. Because there are 5 days (categories), if the frequencies are equal, then each day would have a probability of 20% $\left(\text{or } \frac{1}{5}\right)$.

Hypotheses:

$$\begin{array}{l} H_0: p_1 = p_2 = p_3 = p_4 = p_5 = 20\% \\ H_a: \text{at least one of the } p_i \neq 20\% \end{array}$$

p – value:

From the question, we have $n = 60$ and $k = 5$. Now we need to calculate out the χ^2 -score for the test.

The observed frequency for each category is the number of observations in the sample that fall into that category. This is the information provided in the sample above.

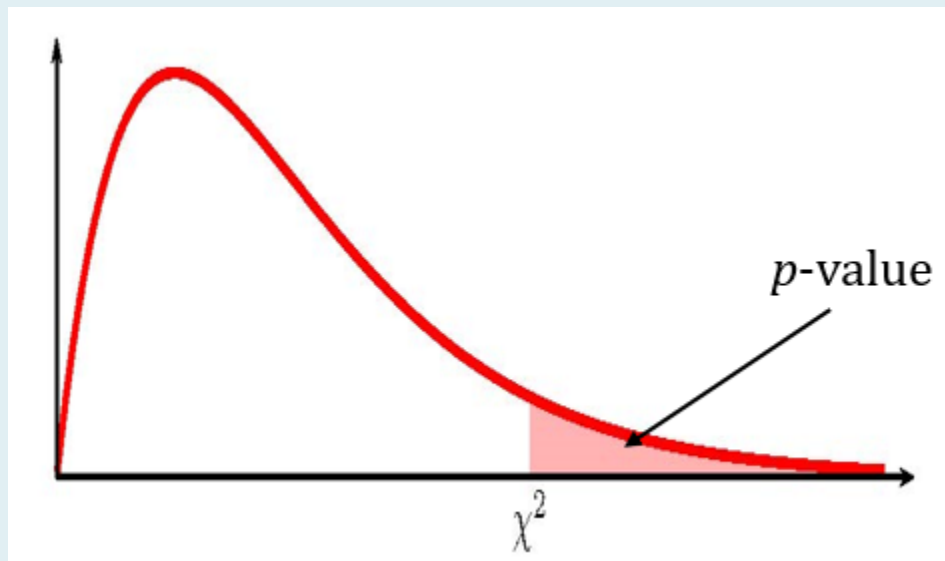
Next, we must calculate out the expected frequencies. The expected frequency is the number of observations we would expect to see in the sample, assuming the null hypothesis is true. To calculate the expected frequency for each category, we multiply the sample size n by the probability associated with that category claimed in the null hypothesis.

Day of the Week	Observed Frequency	Expected Frequency
Monday	15	$0.2 \times 60 = 12$
Tuesday	11	$0.2 \times 60 = 12$
Wednesday	10	$0.2 \times 60 = 12$
Thursday	9	$0.2 \times 60 = 12$
Friday	15	$0.2 \times 60 = 12$

To calculate the χ^2 -score, we work out the quantity $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$ for each category and then add up these quantities.

$$\begin{aligned}
 \chi^2 &= \sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}} \\
 &= \frac{(15-12)^2}{12} + \frac{(11-12)^2}{12} + \frac{(10-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(15-12)^2}{12} \\
 &= 2.666\dots
 \end{aligned}$$

The degrees of freedom for the χ^2 -distribution is $df = k - 1 = 5 - 1 = 4$. The χ^2 -goodness-of-fit test is a right-tailed test, so we use the **chisq.dist.rt** function to find the p -value:



Function	chisq.dist.rt
Field 1	2.666....
Field 2	4
Answer	0.6151

So the p – value = 0.6151.

Conclusion:

Because p – value = 0.6151 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level, there is enough evidence to suggest that the day of the week with the highest number of absences occurs with equal frequency during a five-day work week.

NOTES

1. The null hypothesis is the claim that the probability each day of the week has the highest number of absences is 20%.
2. The alternative hypothesis is the claim that at least one of the probabilities is not 20%. The alternative hypothesis does not say that every p_i does not equal 20%, only that one of them does not equal 20%.
3. Keep all of the decimals throughout the calculation (i.e. in the calculation of the χ^2 -score) to avoid any round-off error in the calculation of the p – value. This ensures that we get the most accurate value for the p – value.
4. The p – value of 0.6151 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis.

TRY IT

Teachers want to know which night each week their students are doing most of their homework. Most teachers think that students do an equal amount of homework each night. Suppose a random sample of 49 students are asked on which night of the week they did the most homework. The results are shown in the table below. At the 5% significance level, are the nights that students do most of their homework equally distributed?

Day of Week	Number of Students
Sunday	11
Monday	8
Tuesday	10
Wednesday	7
Thursday	10
Friday	5
Saturday	5

Click to see Solution

Let p_1 be the probability students do their homework on Sunday, p_2 be the probability students do their homework on Monday, p_3 be the probability students do their homework on Tuesday, p_4 be the probability students do their homework on Wednesday, p_5 be the probability students do their homework on Thursday, p_6 be the probability students do their homework on Friday, and p_7 be the probability students do their homework on Saturday.

Hypotheses:

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = \frac{1}{7}$$

$$H_a : \text{at least one of the } p_i \neq \frac{1}{7}$$

p – value:

From the question, we have $n = 49$ and $k = 7$.

Day of the Week	Observed Frequency	Expected Frequency
Sunday	11	$1/7 \times 49 = 7$
Monday	8	$1/7 \times 49 = 7$
Tuesday	10	$1/7 \times 49 = 7$
Wednesday	7	$1/7 \times 49 = 7$
Thursday	10	$1/7 \times 49 = 7$
Friday	5	$1/7 \times 49 = 7$
Saturday	5	$1/7 \times 49 = 7$

$$\begin{aligned}
 \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\
 &= \frac{(11 - 7)^2}{7} + \frac{(8 - 7)^2}{7} + \frac{(10 - 7)^2}{7} + \frac{(7 - 7)^2}{7} \\
 &\quad + \frac{(10 - 7)^2}{7} + \frac{(5 - 7)^2}{7} + \frac{(5 - 7)^2}{7} \\
 &= 6.142 \dots
 \end{aligned}$$

The degrees of freedom for the χ^2 -distribution is $df = k - 1 = 7 - 1 = 6$.

Function	chisq.dist.rt
Field 1	6.142....
Field 2	6
Answer	0.4074

So the p – value = 0.4074.

Conclusion:

Because p – value = 0.4074 $>$ 0.05 = α , we do not reject the null hypothesis. At the 5% significance level, there is enough evidence to suggest that the nights students do most of their homework are equally distributed.

TRY IT

One study indicates that the number of televisions that American families have is distributed as shown in this table:

Number of Televisions	Percent
0	10\%
1	16\%
2	55\%
3	11\%
4 or more	8\%

A researcher wants to determine if the number of televisions that families in the far western part of the U.S. have the same distribution as the above study. A random sample of 600 families in the far western U.S. is taken, and the results are recorded in the following table:

Number of Televisions	Observed Frequency
0	66
1	119
2	340
3	60
4 or more	15

At the 1\% significance level, does it appear that the distribution of the number of televisions for families in the far western U.S is different from the distribution for the American population as a whole?

Click to see Solution

Let p_1 be the probability a family owns 0 televisions, p_2 be the probability a family owns 1

television, p_3 be the probability a family owns 2 televisions, p_4 be the probability a family owns 3 televisions, and p_5 be the probability a family owns 4 or more televisions.

Hypotheses:

$$\begin{array}{l} H_0: p_1=0.1, p_2=0.16, p_3=0.55, p_4=0.11, p_5=0.08 \\ H_a: \text{at least one of the } p_i \text{ does not equal its stated probability} \end{array}$$

p – value:

From the question, we have $n = 600$ and $k = 5$.

Number of Televisions	Observed Frequency	Expected Frequency
0	66	$0.1 \times 600 = 60$
1	119	$0.16 \times 600 = 96$
2	340	$0.55 \times 600 = 330$
3	60	$0.11 \times 600 = 66$
4 or more	15	$0.08 \times 600 = 48$

$$\begin{aligned} \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(66 - 60)^2}{60} + \frac{(119 - 96)^2}{96} + \frac{(340 - 330)^2}{330} + \frac{(60 - 66)^2}{66} + \frac{(15 - 48)^2}{48} \\ &= 29.646 \dots \end{aligned}$$

The degrees of freedom for the χ^2 -distribution is $df = k - 1 = 5 - 1 = 4$.

Function	chisq.dist.rt
Field 1	29.646....
Field 2	4
Answer	0.000006

So the p – value = 0.000006.

Conclusion:

Because p – value = 0.000006 < 0.01 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 1% significance level, there is enough evidence to suggest that the

distribution of the number of televisions for families in the far western U.S is different from the distribution for the American population as a whole.

TRY IT

The expected percentage of the number of pets students in the United States have in their homes is distributed as follows:

Number of Pets	Percent
0	18\%
1	25\%
2	30\%
3	18\%
4 or more	9\%

A researcher wants to find out if the distribution of the number of pets students in Canada have is the same as the distribution shown in the U.S. A random sample of 1,000 students from Canada is taken, and the results are shown in the table below:

Number of Pets	Observed Frequency
0	210
1	240
2	320
3	140
4+	90

At the 1\% significance level, is the distribution of the number of pets students in Canada have different from the distribution for the United States?

Click to see Solution

Let p_1 be the probability a student owns 0 pets, p_2 be the probability a student owns 1 pet, p_3 be the probability a student owns 2 pets, p_4 be the probability a student owns 3 pets, and p_5 be the probability a student owns 4 or more pets.

Hypotheses:

$$\begin{array}{l} H_0: p_1=0.18, p_2=0.25, p_3=0.30, p_4=0.18, p_5=0.09 \\ H_a: \text{at least one of the } p_i \text{ does not equal its stated probability} \end{array}$$

p – value:

From the question, we have $n = 1000$ and $k = 5$.

Number of Pets	Observed Frequency	Expected Frequency
0	210	$0.18 \times 1000 = 180$
1	240	$0.25 \times 1000 = 250$
2	320	$0.30 \times 1000 = 300$
3	140	$0.18 \times 1000 = 180$
4 or more	90	$0.09 \times 1000 = 90$

$$\begin{aligned} \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(210 - 180)^2}{180} + \frac{(240 - 250)^2}{250} + \frac{(320 - 300)^2}{300} + \frac{(140 - 180)^2}{180} + \frac{(90 - 90)^2}{90} \\ &= 15.622 \dots \end{aligned}$$

The degrees of freedom for the χ^2 -distribution is $df - k - 1 = 5 - 1 = 4$.

Function	chisq.dist.rt
Field 1	15.622....
Field 2	4
Answer	0.0036

So the p – value = 0.0036.

Conclusion:

Because $p\text{-value} = 0.0036 < 0.01 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 1\% significance level, there is enough evidence to suggest that the distribution of the number of pets students in Canada have is different from the distribution for the United States.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=246#oembed-1>

Video: “Pearson’s chi square test (goodness of fit) | Probability and Statistics | Khan Academy” by Khan Academy [11:48] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A teacher predicts that the distribution of grades on the final exam will be as follows:

Grade	Percent
A	25\%
B	30\%
C	35\%
D	10\%

In a class of 20 students, the frequency of the grades on the final exam is given below:

Grade	Frequency
A	7
B	7
C	5
D	1

At the 5% significance level, do the actual grades match the teacher's assumed distribution?

Click to see Answer

- Hypotheses:

$$H_0: p_1=25\%, p_2=30\%, p_3=35\%, p_4=10\%$$

$$H_a: \text{at least one of the } p_i \text{ does not equal its stated probability}$$
- p – value = 0.5645
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the distribution of grades on the final exam follows the teacher's stated distribution.

2. A six-sided die is rolled 120 times, and the results are recorded in the table below. At the 5% significance level, determine if the die is fair. (Hint: in a fair die, each of the faces is equally likely to occur.)

Face Value	Frequency
1	15
2	29
3	16
4	15
5	30
6	15

Click to see Answer

- Hypotheses:

$$H_0: p_1=p_2=p_3=p_4=p_5=p_6=16.67\%$$

$$H_a: \text{at least one of the } p_i \text{ does not equal } 16.67\%$$

- $p - \text{value} = 0.0184$
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that the distribution of the dice rolls does not follow the assumed distribution. The dice is not fair.

3. The distribution of the marital status for the male population of certain country, ages 15 and older, is as shown in the table below.

Marital Status	Percent
never married	31.3\%
married	56.1\%
widowed	2.5\%
divorced/separated	10.1\%

Suppose that a random sample of 400 young adult males from the country, ages 18 to 24 years old, yield the following frequency distribution.

Marital Status	Frequency
never married	140
married	238
widowed	2
divorced/separated	20

At the 1\% significance level, test if this young adult male age group fits the distribution of the adult male population of the country.

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p_1=31.3\%, p_2=56.1\%, p_3=2.5\%, p_4=10.1\% \\ H_a: \text{at least one of the } p_i \text{ does not equal its stated probability} \end{array}$$
- $p - \text{value} = 0.0002$
- Conclusion: At the 1\% significance level, there is enough evidence to conclude that the distribution of the marital statuses of the young adult male age group is different than the

distribution of the adult male population.

4. The columns in the table below contain the Race/Ethnicity of the high schools in a certain country for a recent year and the percentages of the Overall Student Population. A local school district wants to determine if its high schools follow the same distribution for the ethnicity of its students. The school district takes a sample of 1, 000 high school students in the district, and the right column in the table contains the breakdown of the ethnicity of the students in the sample.

Race/Ethnicity	Overall Student Population	Survey Frequency
Asian, Asian American, or Pacific Islander	5.4\%	82
Black	14.5\%	135
Hispanic or Latino	15.9%	136
Indigenous	1.2\%	10
White	61.6\%	604
Not reported/other	1.4\%	33

At the 5\% significance level, determine if the distribution of ethnicity at the local school district follows the overall student population.

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p_1=5.4\%, p_2=14.5\%, p_3=15.9\%, p_4=1.2\%, p_5=61.6\%, p_6=1.4\% \\ H_a: \text{at least one of the } p_i \text{ does not equal its stated probability} \end{array}$$
- p – value = 0.0018
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that the distribution of the ethnicity of high schools in the local district is different than the distribution of the overall high school student population.

5. The table below shows the expected distribution of majors of all male university students across the country. A local university wants to know if the distribution of majors for its male students follows the same distribution. A sample of 5, 000 male students at the university is

taken, and their majors are recorded. The data from the sample is shown in the right column in the table.

Major	Expected Major	Actual Major
Arts & Humanities	12.0\%	630
Biological Sciences	6.7\%	320
Business	22.7\%	1100
Education	5.8\%	315
Engineering	15.6\%	800
Physical Sciences	3.6\%	175
Professional	9.3\%	450
Social Sciences	7.6\%	370
Technical	1.8\%	90
Other	8.2\%	400
Undecided	6.7\%	350

At the 5\% significance level, determine if the distribution of the majors of male students at the local university fits the distribution of majors for all male university students.

Click to see Answer

- Hypotheses:

$$\begin{aligned} H_0: p_1=12\%, p_2=6.7\%, p_3=22.7\%, p_4=5.8\%, p_5=15.6\%, p_6=3.6\%, p_7=9.3\%, p_8=7.6\%, p_9=1.8\%, p_{10}=8.2\%, p_{11}=6.7\% \\ H_a: \text{at least one of the } p_i \text{ does not equal its stated probability} \end{aligned}$$
- p – value = 0.6561
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that the distribution of majors for male students at the local university follows the distribution of majors for all male university students across the country.

6. The table below shows the expected distribution of majors of all female university students across the country. A local university wants to know if the distribution of majors for its female students follows the same distribution. A sample of 5, 000 female students at the university is taken, and their majors are recorded. The data from the sample is shown in the right column

in the table.

Major	Expected Major	Actual Major
Arts & Humanities	14.0\%	670
Biological Sciences	8.4\%	410
Business	13.1\%	685
Education	13.0\%	650
Engineering	2.6\%	145
Physical Sciences	2.6\%	125
Professional	18.9\%	975
Social Sciences	13.0\%	605
Technical	0.4\%	15
Other	5.8\%	300
Undecided	8.2\%	420

At the 5\% significance level, determine if the distribution of majors of female students at the local university fits the distribution of majors for all female university students.

Click to see Answer

- Hypotheses:

$$\begin{array}{l} H_0: p_1=14\%, p_2=8.4\%, p_3=13.1\%, p_4=13\%, p_5=2.6\%, p_6=2.6\%, p_7=18.9\%, p_8=13\%, p_9=0.4\%, p_{10}=5.8\%, p_{11}=8.2\% \\ H_a: \text{at least one of the } p_i \text{ does not equal its stated probability} \end{array}$$

- $p - \text{value} = 0.3791$

- Conclusion: At the 5\% significance level, there is enough evidence to conclude that the distribution of majors for female students at the local university follows the distribution of majors for all female university students across the country.

7. A local police department wants to know if the percentage of traffic accidents is the same for each day of the week. The department takes a sample of 500 traffic accidents and records the day of the week on which they occurred.

Day of the Week	Frequency
Sunday	75
Monday	60
Tuesday	74
Wednesday	57
Thursday	65
Friday	79
Saturday	90

At the 5% significance level, determine if the proportion of traffic accidents is the same for each day of the week.

Click to see Answer

- Hypotheses:

$$H_0: p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = 14.29\%$$

$$H_a: \text{at least one of the } p_i\text{'s does not equal } 14.29\%$$
- p – value = 0.0817
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the proportion of traffic accidents is the same for each day of the week.

8. A local retailer provides a variety of payment options for its customers: cash, cheque, credit and debit. The retailer believes that the current distribution of the payment options is as follows: 15% cash, 5% cheque, 50% credit, and 30% debit. The retailer takes a sample of 450 transactions and records the payment method

Payment Method	Frequency
Cash	78
Cheque	12
Credit	250
Debit	110

At the 1% significance level, determine if the retailer's claimed distribution of the payment methods is correct.

Click to see Answer

- Hypotheses:

$$H_0: p_1 = 15\%, p_2 = 5\%, p_3 = 50\%, p_4 = 30\% \\ H_a: \text{at least one of the } p_i \text{ does not equal its stated probability}$$
- p – value = 0.003
- Conclusion: At the 1% significance level, there is enough evidence to conclude that the distribution of the payment methods is different than the retailer's claim.

9. A local restaurant owner has five locations across the city. The owner wants to know if the percentage of customers is the same at each location. The owner takes a sample of 750 customers and records which restaurant location they visited.

Location	Frequency
1	126
2	179
3	141
4	131
5	173

At the 1% significance level, determine if the proportion of customers is the same for each restaurant location.

Click to see Answer

- Hypotheses:

$$H_0: p_1 = p_2 = p_3 = p_4 = p_5 = 20\% \\ H_a: \text{at least one of the } p_i \text{ does not equal } 20\%$$
- p – value = 0.0031
- Conclusion: At the 1% significance level, there is enough evidence to conclude that the proportion of customers is not the same for each location.

“10.4 The Goodness-of-Fit Test” and “10.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

10.4 THE TEST OF INDEPENDENCE

LEARNING OBJECTIVES

- Conduct and interpret the χ^2 test of independence.

Given two categorical variables, is there some relationship between the two categorical variables, or are the two categorical variables independent? The χ^2 test of independence allows us to test if two categorical variables are independent (not related) or dependent (related). The test of independence can only show if a relationship exists between two variables, but the test does not show if one variable causes changes in the other variable.

The test of independence uses a contingency table to analyze the data. As we saw previously in probability, a contingency table lists the categories of one variable as the rows and the categories of the other variable as the columns. The frequency of a row-column combination is the number of items that occur in both categories.

Conducting a χ^2 Test of Independence

Suppose one categorical variable has r possible outcomes (categories) and the other categorical variable has c possible outcomes (categories).

1. Write down the null and alternative hypotheses:

H_0 : The two categorical variables are independent

H_a : The two categorical variables are dependent

2. Collect the sample information for the test and identify the significance level α .

3. Use the χ^2 -distribution to find the p – value, which is the **area in the right tail** of the distribution. The χ^2 -score and degrees of freedom are

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$df = (r - 1) \times (c - 1)$$

observed = observed frequency from the sample data

$$\text{expected} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

r = number of rows in the contingency table

c = number of columns in the contingency table

4. Compare the p – value to the significance level and state the outcome of the test.
- If p – value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If p – value $> \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
5. Write down a concluding sentence specific to the context of the question.

NOTES

1. The null hypothesis is the claim that the two categorical variables are independent. That is, the null hypothesis claims that there is no relationship between the two categorical variables.
2. The alternative hypothesis is the claim that the two categorical variables are dependent. That is, the alternative hypothesis claims that there is some relationship between the two categorical variables.

3. The test can only show if a relationship exists between the two categorical variables. The test cannot show any type of causal relationship between the two categorical variables.
4. The formula to find the expected frequencies follows from the assumption that the null hypothesis is true and how we calculate joint probabilities for independent events.
Assuming the null hypothesis is true means that we assume the variables are independent. This means that we assume that the events in any row and column combination of the contingency tables are independent. As we saw in probability, when two events A and B are independent, $P(A \text{ and } B) = P(A) \times P(B)$. Using this fact, we get the formula for the expected frequency: $\text{expected} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$.
5. In order to use the χ^2 test of independence, the expected frequency for a cell in the contingency table must be at least 5.
6. The p – value for a χ^2 test of independence is always the area in the right tail of the χ^2 -distribution. So, we use **chisq.dist.rt** to find the p – value for a χ^2 test of independence.
7. To calculate the χ^2 -score:
 - For each of the possible outcomes of the categorical variables, calculate $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$:
 - i. Find the difference between the observed frequency (from the sample) and the expected frequency (from the null hypothesis). The expected frequency of any cell of the contingency table when the null hypothesis is true is:
 $\text{expected} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$
 - ii. Square the difference in step (i).
 - iii. Divide the value found in step (iii) by the expected frequency.
 - Add up the values of $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$ for each of the outcomes.

EXAMPLE

A researcher is studying the relationship between drivers who commit speeding violations and drivers who use cell phones while driving. The researcher took a sample of 755 drivers and obtained the information shown in the table below.

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

At the 5% significance level, is there a relationship between drivers who commit speeding violations and drivers who use cell phones while driving?

Solution

Hypotheses:

H_0 : The two variables are independent

H_a : The two variables are dependent

p – value:

From the question, we have $r = 2$ and $c = 2$. Now we need to calculate out the χ^2 -score for the test.

The observed frequency for each cell is the number of observations in the sample that fall into that cell. This is the information provided in the sample above.

Observed Frequencies (Sample Data)			
	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

Next, we must calculate out the expected frequencies. Because we assume the null hypothesis is true (i.e. the variables are independent), the expected frequency in each cell is

$$\text{expected} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

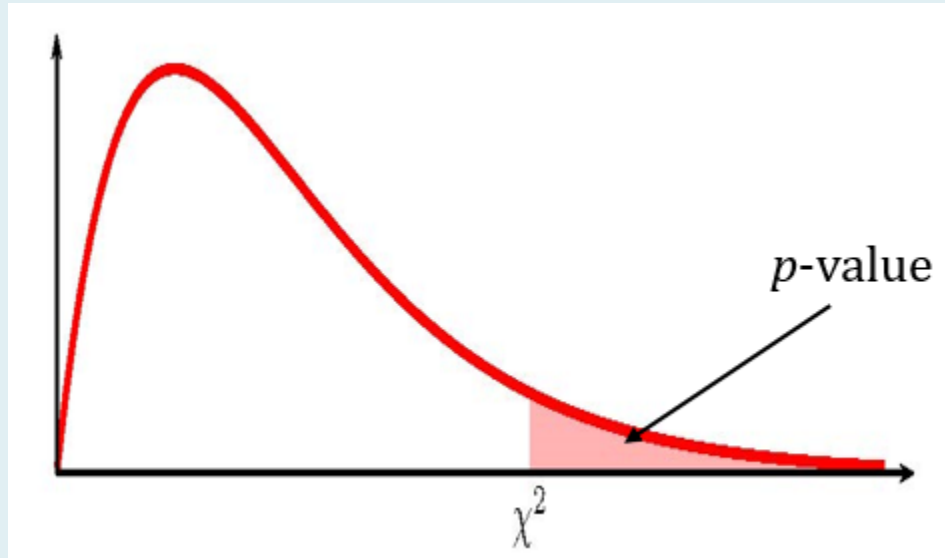
Expected Frequencies			
	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	$\frac{305 \times 70}{755} = 28.27 \dots$	$\frac{305 \times 685}{755} = 276.72 \dots$	305
Not a cell phone user	$\frac{450 \times 70}{755} = 41.72 \dots$	$\frac{450 \times 685}{755} = 408.27 \dots$	450
Total	70	685	755

To calculate the χ^2 -score for each cell, we work out the quantity $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$ and then add up these quantities.

$$\begin{aligned}
 \chi^2 &= \sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}} \\
 &= \frac{(25 - 28.27 \dots)^2}{28.27 \dots} + \frac{(280 - 276.72 \dots)^2}{276.72 \dots} + \frac{(45 - 41.72 \dots)^2}{41.72 \dots} + \frac{(405 - 408.27 \dots)^2}{408.27 \dots} \\
 &= 0.7027 \dots
 \end{aligned}$$

The degrees of freedom for the χ^2 -distribution is

$df = (r - 1) \times (c - 1) = (2 - 1) \times (2 - 1) = 1$. The χ^2 test of independence is a right-tailed test, so we use **chisq.dist.rt** function to find the p - value:



Function	chisq.dist.rt
Field 1	0.7027....
Field 2	1
Answer	0.4019

So the p – value = 0.4019.

Conclusion:

Because p – value = 0.4019 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level, there is enough evidence to suggest that the two variables are independent.

NOTES

1. The null hypothesis is the claim that the variables are independent. That is, there is no relationship between drivers who commit speeding violations and drivers who use cell phones while driving.
2. The alternative hypothesis is the claim that the variables are dependent. That is, there is a relationship between drivers who commit speeding violations and drivers who use cell phones while driving.
3. Keep all of the decimals throughout the calculation (i.e. in the calculation of the χ^2 -score) to

avoid any round-off error in the calculation of the p — value. This ensures that we get the most accurate value for the p — value.

4. The p — value is the area in the right tail of the χ^2 -distribution, to the right of $\chi^2 = 0.7027 \dots$. In the calculation of the p — value:
 - The function is **chisq.dist.rt** because we are finding the area in the right tail of a χ^2 -distribution.
 - Field 1 is the value of χ^2 .
 - Field 2 is the value of the degrees of freedom df .
5. The p — value of **0.4019** is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the two variables are independent.

EXAMPLE

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among college students, university students, and non-students. The table below is a sample of the adult volunteers and the number of hours they volunteer per week.

	1-3 Hours	4-6 Hours	7-9 Hours	Total
College Students	111	96	48	255
University Students	96	133	61	290
Non Students	91	150	53	294
Total	298	379	162	839

At the 5% significance level, is the number of hours volunteered independent of the type of volunteer?

Solution

Hypotheses:

H_0 : The two variables are independent

H_a : The two variables are dependent

p – value:

From the question, we have $r = 3$ and $c = 3$. Now we need to calculate out the χ^2 -score for the test.

The observed frequency for each cell is the number of observations in the sample that fall into that cell. This is the information provided in the sample above.

Observed Frequencies (Sample Data)				
	1-3 Hours	4-6 Hours	7-9 Hours	Total
College Students	111	96	48	255
University Students	96	133	61	290
Non Students	91	150	53	294
Total	298	379	162	839

Next, we must calculate out the expected frequencies. Because we assume the null hypothesis is true (i.e. the variables are independent), the expected frequency in each cell is

$$\text{expected} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

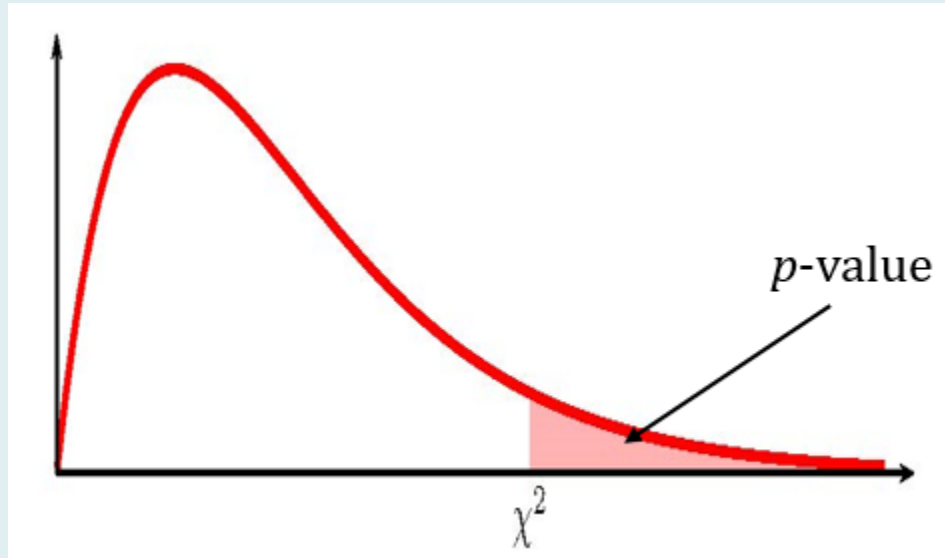
Expected Frequencies				
	1-3 Hours	4-6 Hours	7-9 Hours	Total
College Students	$\frac{255 \times 298}{839} = 90.57 \dots$	$\frac{255 \times 379}{839} = 115.19 \dots$	$\frac{255 \times 162}{839} = 49.23 \dots$	255
University Students	$\frac{290 \times 298}{839} = 103.00 \dots$	$\frac{290 \times 379}{839} = 131.00 \dots$	$\frac{290 \times 162}{839} = 55.99 \dots$	290
Non Students	$\frac{294 \times 298}{839} = 104.42 \dots$	$\frac{294 \times 379}{839} = 132.80 \dots$	$\frac{294 \times 162}{839} = 56.76 \dots$	294
Total	298	379	162	839

To calculate the χ^2 -score for each cell, we work out the quantity $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$ and then add up these quantities.

$$\begin{aligned}
 \chi^2 &= \sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}} \\
 &= \frac{(111 - 90.57 \dots)^2}{90.57 \dots} + \frac{(96 - 115.19 \dots)^2}{115.19 \dots} + \frac{(48 - 49.23 \dots)^2}{49.23 \dots} \\
 &\quad + \frac{(96 - 103.00 \dots)^2}{103.00 \dots} + \frac{(133 - 131.00 \dots)^2}{131.00 \dots} + \frac{(61 - 55.99 \dots)^2}{55.99 \dots} \\
 &\quad + \frac{(91 - 104.42 \dots)^2}{104.42 \dots} + \frac{(150 - 132.80 \dots)^2}{132.80 \dots} + \frac{(53 - 56.76 \dots)^2}{56.76 \dots} \\
 &= 12.99 \dots
 \end{aligned}$$

The degrees of freedom for the χ^2 -distribution is

$df = (r - 1) \times (c - 1) = (3 - 1) \times (3 - 1) = 4$. The χ^2 test of independence is a right-tailed test, so we use **chisq.dist.rt** function to find the p - value:



Function	chisq.dist.rt
Field 1	12.99....
Field 2	4
Answer	0.0113

So the p – value = 0.0113.

Conclusion:

Because p – value = 0.0113 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5\% significance level, there is enough evidence to suggest that the number of hours volunteered and the type of volunteer are dependent.

NOTES

1. The null hypothesis is the claim that the variables are independent. That is, there is no relationship between the number of hours volunteered and the type of volunteer.
2. The alternative hypothesis is the claim that the variables are dependent. That is, there is a relationship between the number of hours volunteered and the type of volunteer.
3. Keep all of the decimals throughout the calculation (i.e. in the calculation of the χ^2 -score) to avoid any round-off error in the calculation of the p – value. This ensures that we get the

most accurate value for the p — value.

4. The p — value of **0.0113** is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the two variables are dependent.

TRY IT

In a local school district, a music teacher wants to study the relationship between students who take music and students on the honour roll. The teacher took a sample of **300** students and obtained the information shown in the table below.

	Honour Roll Student	Non-Honour Roll Student	Total
Music Student	24	26	50
Non-Music Student	67	183	250
Total	97	203	300

At the 5% significance level, is there a relationship between music/non-music students and honour roll/non-honour roll students?

Click to see Solution

Hypotheses:

H_0 : The two variables are independent

H_a : The two variables are dependent

p – value:

From the question, we have $r = 2$ and $c = 2$.

Observed Frequencies (Sample Data)			
	Honour Roll Student	Non-Honour Roll Student	Total
Music Student	24	26	50
Non-Music Student	67	183	250
Total	97	203	300

Expected Frequencies			
	Honour Roll Student	Non-Honour Roll Student	Total
Music Student	16.166 ...	33.833 ...	50
Non-Music Student	80.833 ...	169.166 ...	250
Total	97	203	300

$$\begin{aligned}
 \chi^2 &= \sum \frac{(\text{observed-expected})^2}{\text{expected}} \\
 &= \frac{(24 - 16.166 \dots)^2}{16.166 \dots} + \frac{(26 - 33.833 \dots)^2}{33.833 \dots} + \frac{(67 - 80.833 \dots)^2}{80.833 \dots} + \frac{(183 - 169.166 \dots)^2}{169.165 \dots} \\
 &= 9.107 \dots
 \end{aligned}$$

$$\begin{aligned}
 df &= (r - 1) \times (c - 1) \\
 &= (2 - 1) \times (2 - 1) \\
 &= 1
 \end{aligned}$$

Function	chisq.dist.rt
Field 1	9.107...
Field 2	1
Answer	0.0025

So the p – value = 0.0025.

Conclusion:

Because p – value = 0.0025 < 0.05 = α , we reject the null hypothesis in favour of the

alternative hypothesis. At the 5\% significance level, there is enough evidence to suggest that the two variables are dependent.

TRY IT

A local college is interested in the relationship between student anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and the need to succeed in school. The results are shown in the table below.

	High Anxiety	Med-High Anxiety	Medium Anxiety	Med-Low Anxiety	Low Anxiety	Total
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Total	57	95	127	63	58	400

At the 5\% significance level, is there a relationship between student anxiety level and the need to succeed in school?

Click to see Solution

Hypotheses:

H_0 : The two variables are independent

H_a : The two variables are dependent

p – value:

From the question, we have $r = 3$ and $c = 5$.

Observed Frequencies (Sample Data)						
	High Anxiety	Med-High Anxiety	Medium Anxiety	Med-Low Anxiety	Low Anxiety	Total
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Total	57	95	127	63	58	400

Expected Frequencies						
	High Anxiety	Med-High Anxiety	Medium Anxiety	Med-Low Anxiety	Low Anxiety	Total
High Need	22.08 ...	36.81 ...	49.21 ...	24.41 ...	22.47 ...	155
Medium Need	27.50 ...	45.83 ...	61.27 ...	30.39 ...	27.98 ...	193
Low Need	7.41	12.35	16.51	8.19	7.54	52
Total	57	95	127	63	58	400

$$\begin{aligned}
 \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\
 &= \frac{(35 - 22.08 \dots)^2}{22.08 \dots} + \frac{(18 - 27.50 \dots)^2}{27.50 \dots} + \frac{(4 - 7.41)^2}{7.41} \\
 &\quad + \frac{(42 - 36.81 \dots)^2}{36.81 \dots} + \frac{(48 - 45.83 \dots)^2}{45.83 \dots} + \frac{(5 - 12.35)^2}{12.35} \\
 &\quad + \frac{(53 - 49.21 \dots)^2}{49.21 \dots} + \frac{(63 - 61.27 \dots)^2}{61.27 \dots} + \frac{(11 - 16.51)^2}{16.51} \\
 &\quad + \frac{(15 - 24.41 \dots)^2}{24.41 \dots} + \frac{(33 - 30.39 \dots)^2}{30.39 \dots} + \frac{(15 - 8.19)^2}{8.19} \\
 &\quad + \frac{(10 - 22.47 \dots)^2}{22.47 \dots} + \frac{(31 - 27.98 \dots)^2}{27.98 \dots} + \frac{(17 - 7.54)^2}{7.54} \\
 &= 48.419 \dots
 \end{aligned}$$

$$\begin{aligned}
 df &= (r - 1) \times (c - 1) \\
 &= (3 - 1) \times (5 - 1) \\
 &= 8
 \end{aligned}$$

Function	chisq.dist.rt
Field 1	48.419...
Field 2	8
Answer	0.00000008

So the p – value = 0.00000008.

Conclusion:

Because p – value = 0.00000008 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level, there is enough evidence to suggest that the two variables are dependent.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=248#oembed-1>

Video: “Chi-square test for association (independence) | AP Statistics | Khan Academy” by Khan Academy [10:28] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. Transit Railroads is interested in the relationship between travel distance and the ticket class purchased. A random sample of 200 passengers is taken. The table below shows the results. At the 5% significance level, determine if a passenger’s choice in ticket class is independent of the distance they must travel.

Travelling Distance	Third class	Second class	First class	Total
1–100 km	21	14	6	41
101–200 km	18	16	8	42
201–300 km	16	17	15	48
301–400 km	12	14	21	47
401–500 km	6	6	10	22
Total	73	67	60	200

Click to see Answer

- Hypotheses: H_0 : The two variables are independent
 H_a : The two variables are dependent
- p – value = 0.0435
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that a passenger's choice of ticket and distance traveled are dependent.

2. A recent debate about where in Canada skiers believe that the skiing is best prompted the following survey. At the 5\% significance level test to see if the best ski area is independent of the level of the skier.

Ski Area	Beginner	Intermediate	Advanced	Total
B.C.	20	30	40	90
Alberta	10	30	60	100
Quebec	10	40	50	100
Total	40	100	150	290

Click to see Answer

- Hypotheses: H_0 : The two variables are independent
 H_a : The two variables are dependent
- p – value = 0.0324
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that the best ski area and the level of the skier are dependent.

3. Car manufacturers are interested in whether there is a relationship between the size of the car an individual drives and the number of people in the driver's family (that is, whether car size and family size are independent). To test this, suppose that 800 car owners were randomly surveyed with the results in the table. Conduct a test of independence. Use a 5% significance level.

Family Size	Sub & Compact	Mid-size	Full-size	Van & Truck	Total
1	20	35	40	35	130
2	20	50	70	80	220
3-4	20	50	100	90	260
5+	20	30	70	70	190
Total	80	165	280	275	800

Click to see Answer

- Hypotheses: H_0 : The two variables are independent
 H_a : The two variables are dependent
- p – value = 0.1581
- Conclusion: At the 5% significance level, there is enough evidence to conclude that car size and family size are independent.

4. College students may be interested in whether or not their majors have any effect on starting salaries after graduation. Suppose that 300 recent graduates were surveyed as to their majors in college and their starting salaries after graduation. The table below shows the data. At the 1% significance level, test if college major and starting salaries are independent.

Major	Less than \$50,000	\$50,000 – \$68,999	More than \$69,000	Total
English	5	20	5	30
Engineering	10	30	60	100
Nursing	10	15	15	40
Business	10	20	30	60
Psychology	20	30	20	70
Total	55	115	130	300

Click to see Answer

- Hypotheses: H_0 : The two variables are independent
 H_a : The two variables are dependent
- p – value = 0.0012
- Conclusion: At the 1\% significance level, there is enough evidence to conclude that college majors and starting salaries are dependent.

5. Some travel agents claim that honeymoon hot spots vary according to the age of the bride. Suppose that 280 recent brides were interviewed as to where they spent their honeymoons. The information is recorded in the table below. At the 5\% significance level, test if the honeymoon location and age of the bride are independent.

Location	20–29	30–39	40–49	50 and over
Niagara Falls	15	25	25	20
Poconos	15	25	25	10
Europe	10	25	15	5
Virgin Islands	20	25	15	5

Click to see Answer

- Hypotheses: H_0 : The two variables are independent
 H_a : The two variables are dependent
- p – value = 0.0734
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that the

honeymoon location and age of the bride are independent.

6. A manager of a sports club keeps information concerning the main sport in which members participate and their ages. To test whether there is a relationship between the age of a member and his or her choice of sport, 643 members of the sports club are randomly selected. At the 5% significance level, test if the age of the member and sport of choice are independent.

Sport	18 – 25	26 – 30	31 – 40	41 and over
racquetball	42	58	30	46
tennis	58	76	38	65
swimming	72	60	65	33

Click to see Answer

- Hypotheses: H_0 : The two variables are independent
 H_a : The two variables are dependent
- p – value = 0.0003
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the age of the member and sport of choice is dependent.

7. A major food manufacturer is concerned that the sales of its skinny french fries have been decreasing. As a part of a feasibility study, the company conducts research into the types of fries sold across the country to determine if the type of fries sold is independent of the area of the country. The results of the study are shown in the table. Conduct a test of independence. Use a 5% significance level.

Type of Fries	Northeast	South	Central	West	Total
skinny fries	70	50	20	25	165
curly fries	100	60	15	30	205
steak fries	20	40	10	10	80
Total	190	150	45	65	450

Click to see Answer

- Hypotheses: H_0 : The two variables are independent
 H_a : The two variables are dependent
- p – value = 0.0044
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that the type of fries and area of the country are dependent.

8. According to a local insurance agent, the following is a breakdown of the amount of life insurance purchased by males in the following age groups. The insurance agent is interested in whether the age of the male and the amount of life insurance purchased are independent events. At a 5\% significance level, test if the age of males and the amount of life insurance purchased are independent.

Age of Males	None	Less than \$200,000	\$200,000–\$400,000	\$401,001–\$1,000,000	More than \$1,000,001
20–29	40	15	40	10	15
30–39	35	16	20	16	10
40–49	20	35	30	22	10
50+	40	30	15	15	10

Click to see Answer

- Hypotheses: H_0 : The two variables are independent
 H_a : The two variables are dependent
 - p – value = 0.0003
 - Conclusion: At the 5\% significance level, there is enough evidence to conclude that the age of males and the amount of life insurance purchased are dependent.
9. Suppose that 600 thirty-year-olds were surveyed to determine whether or not there is a relationship between a person's level of education and salary. At the 5\% significance level, test if a person's level of education and salary are independent.

Annual Salary	Not a high school graduate	High school graduate	College Graduate	Masters or doctorate
Less than \$30,000	10	15	15	25
\$30,000–\$40,000	15	30	45	60
\$40,000–\$50,000	10	20	40	55
\$50,000–\$60,000	5	15	20	60
More than \$60,000	10	20	30	100

Click to see Answer

- Hypotheses: H_0 : The two variables are independent
 H_a : The two variables are dependent
- p – value = 0.0018
- Conclusion: At the 5% significance level, there is enough evidence to conclude that a person's level of education and salary are dependent.

10. An ice cream maker performs a nationwide survey about favourite flavours of ice cream in different geographic areas of the U.S. The results of the survey are given in the table below. At the 5% significance level, test if geographic location and favourite flavour of ice cream are independent.

U.S. region/ Flavor	Strawberry	Chocolate	Vanilla	Rocky Road	Mint Chocolate Chip	Pistachio	Total
West	12	21	22	19	15	8	97
Midwest	10	32	22	11	15	6	96
East	8	31	27	8	15	7	96
South	15	28	30	8	15	6	102
Total	45	112	101	46	60	27	391

Click to see Answer

- Hypotheses: H_0 : The two variables are independent
 H_a : The two variables are dependent

- p – value = 0.5207
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that location and favourite flavour of ice cream are independent.

11. The table provides a recent survey of the youngest online entrepreneurs whose net worth is estimated at one million dollars or more. Their ages range from 17 to 30. Each cell in the table illustrates the number of entrepreneurs who correspond to the specific age group and their net worth. At the 5\% significance level, are the ages and net worth independent?

Age Group\ Net Worth Value (in millions of US dollars)	1–5	6–24	25 or older	Total
17–25	8	7	5	20
26–30	6	5	9	20
Total	14	12	14	40

Click to see Answer

- Hypotheses: H_0 : The two variables are independent
 H_a : The two variables are dependent
- p – value = 0.4144
- Conclusion: At the 5\% significance level, there is enough evidence to conclude that ages and net worth are independent.

“10.5 The Test of Independence” and “10.6 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

PART XI

STATISTICAL INFERENCE USING THE F-DISTRIBUTION

Previously, we looked at confidence intervals and hypothesis tests involving two population means and two population proportions, which allowed us to compare the means or proportions between two populations. But what about confidence intervals and hypothesis tests for two population variances? Sometimes, we need to compare the variability between two populations. For example, a manufacturing process might use two different methods to produce a certain item, and we want to know which method has the smaller variability. In our examination of statistical inference for the two population mean and two population proportion, we looked at the difference between the two population parameters. However, in order to compare the variance between two populations, we have to look at the ratio of the variances in order to use the F -distribution.

Another situation that uses the F -distribution is hypothesis testing on the equality of three or more population means. Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the mean amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compare the mean gas mileage of several models. Each of these scenarios requires a hypothesis test on three or more population means. Hypothesis tests that compare means between more than two groups employ a method using the F -distribution called **analysis of variance** (abbreviated ANOVA). The simplest form of ANOVA, called single factor or one-way ANOVA, is the type used to test the equality of three or more population means.

CHAPTER OUTLINE

11.1 The F-Distribution

11.2 Statistical Inference for Two Population Variances

11.3 One-Way ANOVA and Hypothesis Tests for Three or More Population Means

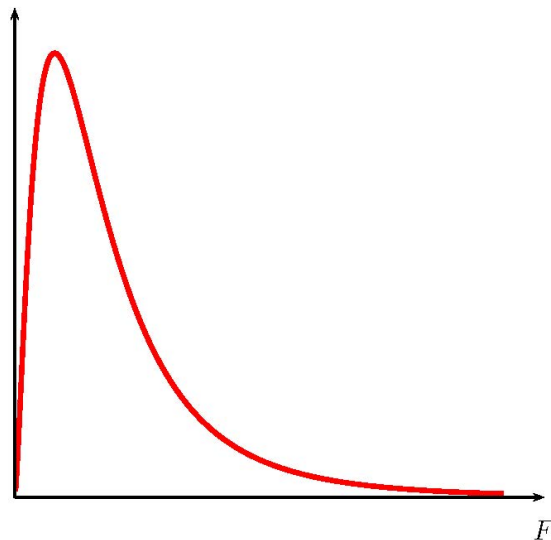
“11.1 Introduction to Statistical Inferences Using the F-Distribution” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

11.1 THE F-DISTRIBUTION

LEARNING OBJECTIVES

- Find the area under an F -distribution.
- Find the F -score for a given area under the curve of an F -distribution.

The F -distribution is a continuous probability distribution. The graph of an F -distribution is shown below. The F -distribution is used in statistical inference to test the equality of population variances, test the equality of three or more population means, or test the overall multiple regression model.



Properties of the F -distribution:

- The graph of an F -distribution is positively skewed and asymmetrical with a minimum value

of 0 and no maximum value.

- An F -distribution is determined by two different degrees of freedom, df_1 and df_2 . df_1 is the degrees of freedom for the numerator of the F -score, and df_2 is the degrees of freedom for the denominator of the F -score. The values of the degrees of freedom depends on how the F -distribution is used. There is a different F -distribution for every set of degrees of freedom. As the values of df_1 and df_2 get larger, the F -distribution approaches a normal distribution.
- The total area under the graph of an F -distribution is 1.
- Probabilities associated with an F -distribution are given by the area under the curve of the F -distribution.

USING EXCEL TO CALCULATE THE AREA UNDER AN F -DISTRIBUTION

To find the area in the left tail:

- To find the area under an F -distribution to the left of a given F -score, use the **f.dist(F , degrees of freedom 1, degrees of freedom 2, logic operator)** function.
 - For F , enter the F -score.
 - For **degrees of freedom 1**, enter the value of df_1 for the F -distribution. df_1 is the degrees of freedom for the numerator of the F -score.
 - For **degrees of freedom 2**, enter the value of df_2 for the F -distribution. df_2 is the degrees of freedom for the denominator of the F -score.
 - For **logic operator**, enter **true**.
- The output from the **f.dist** function is the area to the left of the entered F -score.
- Visit the Microsoft page for more information about the **f.dist** function.

To find the area in the right tail:

- To find the area under an F -distribution to the right of a given F -score, use the **f.dist.rt(F , degrees of freedom 1, degrees of freedom 2)** function.
 - For F , enter the F -score.

- For **degrees of freedom 1**, enter the value of df_1 for the F -distribution. df_1 is the degrees of freedom for the numerator of the F -score.
- For **degrees of freedom 2**, enter the value of df_2 for the F -distribution. df_2 is the degrees of freedom for the denominator of the F -score.
- The output from the **f.dist.rt** function is the area to the right of the entered F -score.
- Visit the Microsoft page for more information about the **f.dist.rt** function.

EXAMPLE

Consider an F -distribution with $df_1 = 12$ and $df_2 = 27$.

1. Find the area under the F -distribution to the left of $F = 0.69$.
2. Find the area under the F -distribution to the right of $F = 1.53$.

Solution

1.	Function	f.dist
	Field 1	0.69
	Field 2	12
	Field 3	27
	Field 4	true
	Answer	0.2535

2.	Function	f.dist.rt
	Field 1	1.53
	Field 2	12
	Field 3	27
	Answer	0.1738

USING EXCEL TO CALCULATE F -SCORES

To find the F -score for the a given left-tail area:

- To find the F -score for a given area under an F -distribution to the left of the F -score, use the **f.inv(area to the left, degrees of freedom 1, degrees freedom 2)** function.
 - For **area to the left**, enter the area to the left of the required F -score.
 - For **degrees of freedom 1**, enter the value of df_1 for the F -distribution. df_1 is the degrees of freedom for the numerator of the F -score.
 - For **degrees of freedom 2**, enter the value of df_2 for the F -distribution. df_2 is the degrees of freedom for the denominator of the F -score.
- The output from the **f.inv** function is the value of the F -score so that the area to the left of the F -score is the entered area.
- Visit the Microsoft page for more information about the **f.inv** function.

To find the F -score for the a given right-tail area:

- To find the F -score for a given area under an F -distribution to the right of the F -score, use the **f.inv.rt(area to the right, degrees of freedom 1, degrees of freedom 2)** function.
 - For **area to the right**, enter the area to the right of required F -score.
 - For **degrees of freedom 1**, enter the value of df_1 for the F -distribution. df_1 is the degrees of freedom for the numerator of the F -score.
 - For **degrees of freedom 2**, enter the value of df_2 for the F -distribution. df_2 is the degrees of freedom for the denominator of the F -score.
- The output from the **f.inv.rt** function is the value of the F -score so that the area to the right of the F -score is the entered area.
- Visit the Microsoft page for more information about the **f.inv.rt** function.

EXAMPLE

Consider an F -distribution with $df_1 = 37$ and $df_2 = 15$.

1. Find the F -score so that the area under the F -distribution to the left of F is 0.413.
2. Find the F -score so that the area under the F -distribution to the right of F is 0.148.

Solution

1.	Function	f.inv
	Field 1	0.413
	Field 2	37
	Field 3	15
	Answer	0.934

2.	Function	f.dist.rt
	Field 1	0.269
	Field 2	37
	Field 3	15
	Answer	1.354

TRY IT

Consider an F -distribution with $df_1 = 17$ and $df_2 = 28$.

1. Find the area under the F -distribution to the right of $F = 2.15$.
2. Find the F -score so that the area under the F -distribution to the left of F is 0.486.
3. Find the F -score so that the area under the F -distribution to the right of F is 0.1493.
4. Find the area under the F -distribution to the left of $F = 1.52$.

Click to see Solution

1.	Function	f.dist.rt
	Field 1	2.15
	Field 2	17
	Field 3	28
	Answer	0.0351

2.	Function	f.inv
	Field 1	0.486
	Field 2	17
	Field 3	28
	Answer	0.969

3.	Function	f.inv.rt
	Field 1	0.1493
	Field 2	17
	Field 3	28
	Answer	1.546

4.	Function	f.dist
	Field 1	1.52
	Field 2	17
	Field 3	28
	Field 4	true
	Answer	0.8415

Exercises

1. Consider an F -distribution with $df_1 = 13$ and $df_2 = 5$.
 - a. Find the area under the F -distribution to the left of $F = 2.79$.
 - b. Find the area under the F -distribution to the right of $F = 4.36$.
 - c. Find the F -score so that the area under the F -distribution to the left of F is 0.2789.
 - d. Find the F -score so that the area under the F -distribution to the right of F is 0.416.

Click to see Answer

- a. 0.8678
 - b. 0.0570
 - c. 0.701
 - d. 1.289
2. Consider a F -distribution with $df_1 = 9$ and $df_2 = 21$.
 - a. Find the area under the F -distribution to the left of $F = 1.07$.
 - b. Find the area under the F -distribution to the right of $F = 3.65$.
 - c. Find the F -score so that the area under the F -distribution to the left of F is 0.6835.
 - d. Find the F -score so that the area under the F -distribution to the right of F is 0.2174.

Click to see Answer

- a. 0.5770
 - b. 0.0069
 - c. 1.255
 - d. 1.484

“11.2 The F-Distribution” and “11.5 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

11.2 STATISTICAL INFERENCE FOR TWO POPULATION VARIANCES

LEARNING OBJECTIVES

- Construct and interpret a confidence interval for two population variances.
- Conduct and interpret a hypothesis test for two population variances.

Sometimes, we want to compare the variability between two populations, instead of comparing the means of the populations. For example, college administrators would like two college professors grading exams to have the same variation in their grading, or a supermarket might be interested in the variability of the check-out times for two checkers.

As with comparing other population parameters, we can construct confidence intervals and conduct hypothesis tests to study the relationship between two population variances. However, because of the distribution we need to use, we study the **ratio of two population variances**, not the difference in the variances.

Throughout this section, we will use subscripts to identify the values for the sample sizes, variances, and standard deviations for the two populations.

Symbol for:	Population 1	Population 2
Population Variance	σ_1^2	σ_2^2
Population Standard Deviation	σ_1	σ_2
Sample Size	n_1	n_2
Sample Variance	s_1^2	s_2^2
Sample Standard Deviation	s_1	s_2

In order to construct a confidence interval or conduct a hypothesis test on the ratio of two population variances, $\frac{\sigma_1^2}{\sigma_2^2}$, we need to use the distribution of $\frac{s_1^2}{s_2^2}$ when the population variances are equal ($\sigma_1^2 = \sigma_2^2$). Suppose we have two normal populations with equal variances $\sigma_1^2 = \sigma_2^2$. A sample of size n_1 with sample variance s_1^2 is taken from population 1 and a sample of size n_2 with sample variance s_2^2 is taken from population 2. The sampling distribution of the ratio of the sample variances $\frac{s_1^2}{s_2^2}$ follows an F -distribution with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.

Constructing a Confidence Interval for the Ratio of Two Population Variances

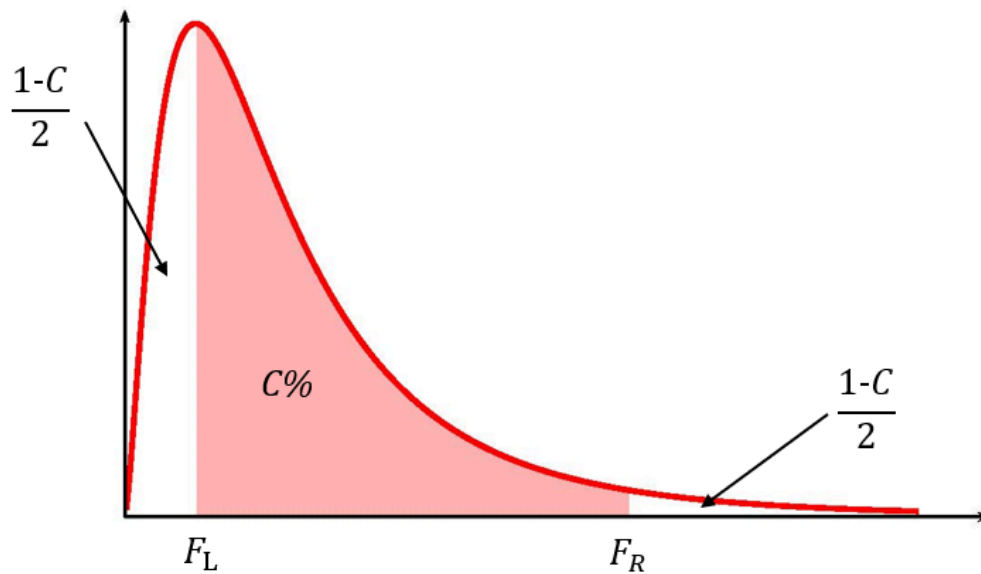
Suppose a sample of size n_1 with sample variance s_1^2 is taken from population 1 and a sample of size n_2 with sample variance s_2^2 is taken from population 2, where the populations are independent and normally distributed. The limits for the confidence interval with confidence level C for the ratio of the population variances $\frac{\sigma_1^2}{\sigma_2^2}$ are

$$\text{Lower Limit} = \frac{1}{F_R} \times \frac{s_1^2}{s_2^2}$$

$$\text{Upper Limit} = \frac{1}{F_L} \times \frac{s_1^2}{s_2^2}$$

where F_L is the F -score so that the area in the left tail of the F -distribution is $\frac{1-C}{2}$, F_R is the

F -score so that the area in the right tail of the F -distribution is $\frac{1 - C}{2}$ and the F -distribution has degrees of freedom $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.



NOTES

1. Like the other confidence intervals we have seen, the F -scores are the values that trap $C\%$ of the observations in the middle of the distribution so that the area of each tail is $\frac{1 - C}{2}$.
2. Because the F -distribution is not symmetrical, the confidence interval for the ratio of the population variances requires that we calculate two different F -scores: one for the left tail and one for the right tail. In Excel, we need to use both the **f.inv** function (for the left tail) and the **f.inv.rt** function (for the right tail) to find the two different F -scores.
3. The F -score for the left tail is part of the formula for the upper limit, and the F -score for the right tail is part of the formula for the lower limit. **This is not a mistake.** It follows from the formula used to determine the limits for the confidence interval.
4. It is important that the populations are independent and normally distributed. If the populations are not normal, the confidence interval will not give an accurate result.

EXAMPLE

Two local walk-in medical clinics want to determine if there is any variability in the time patients wait to see a doctor at each clinic. In a sample of 30 patients at Clinic 1, the standard deviation for the wait time to see a doctor was 45 minutes. In a sample of 40 patients at Clinic 2, the standard deviation for the wait time to see a doctor was 27 minutes. Assume the population of wait times at the two clinics are independent and normally distributed.

1. Construct a 95% confidence interval for the ratio of the variances for the wait times at the two clinics.
2. Interpret the confidence interval found in part 1.
3. Is there evidence to suggest that there is a difference in the variances of the wait times at the two clinics? Explain.

Solution

1. Let Clinic 1 be population 1 and Clinic 2 be population 2. From the question, we have the following information:

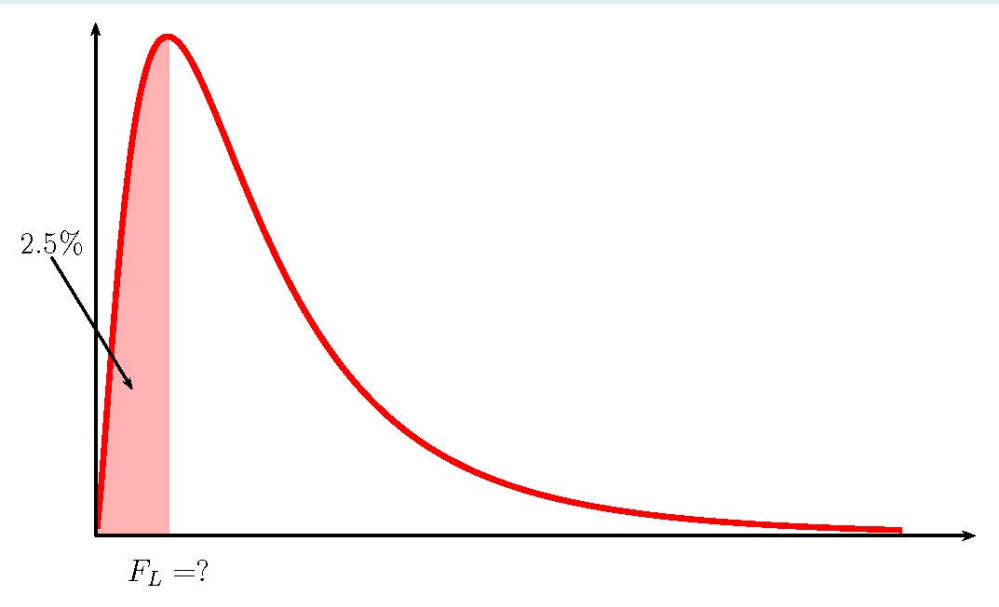
Clinic 1	Clinic 2
$n_1 = 30$	$n_2 = 40$
$s_1^2 = 45^2 = 2025$	$s_2^2 = 27^2 = 729$

To find the confidence interval, we need to find the F_L -score for the 95% confidence interval.

This means that we need to find the F_L -score so that the area in the left tail is

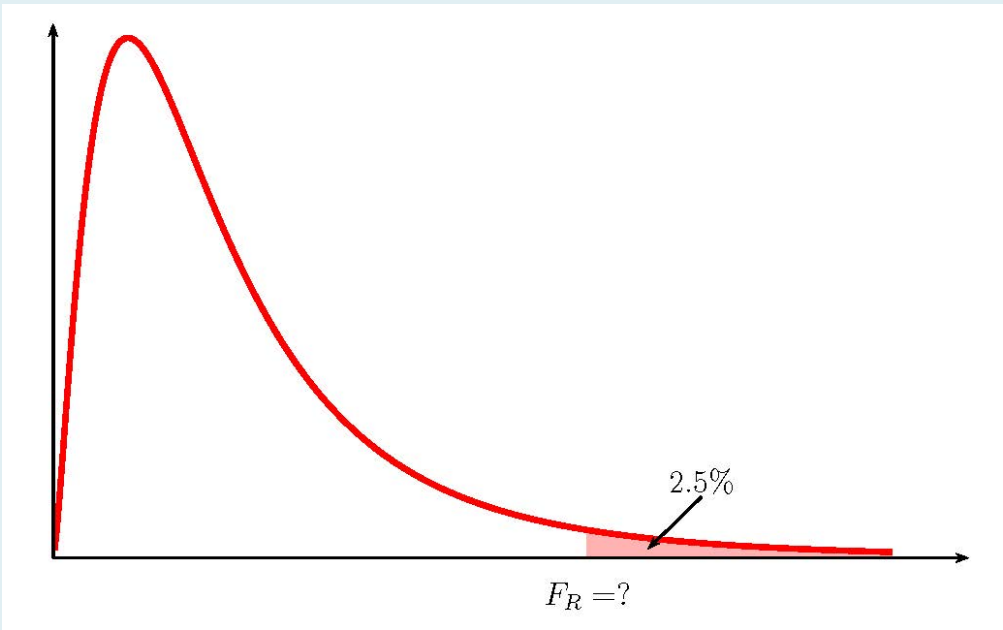
$$\frac{1 - 0.95}{2} = 0.025. \text{ The degrees of freedom for the } F\text{-distribution are}$$

$$df_1 = n_1 - 1 = 30 - 1 = 29 \text{ and } df_2 = n_2 - 1 = 40 - 1 = 39.$$



Function	f.inv
Field 1	0.025
Field 2	29
Field 3	39
Answer	0.4919...

We also need to find the F_R -score for the 95\% confidence interval. This means that we need to find the F_R -score so that the area in the right tail is $\frac{1 - 0.95}{2} = 0.025$. The degrees of freedom for the F -distribution are $df_1 = n_1 - 1 = 30 - 1 = 29$ and $df_2 = n_2 - 1 = 40 - 1 = 39$.



Function	f.inv.rt
Field 1	0.025
Field 2	29
Field 3	39
Answer	1.9618...

So $F_L = 0.4919 \dots$ and $F_R = 1.9618 \dots$. The 95\% confidence interval is

$$\begin{aligned}
 \text{Lower Limit} &= \frac{1}{F_R} \times \frac{s_1^2}{s_2^2} \\
 &= \frac{1}{1.9618\dots} \times \frac{2025}{729} \\
 &= 1.416
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \frac{1}{F_L} \times \frac{s_1^2}{s_2^2} \\
 &= \frac{1}{0.4919\dots} \times \frac{2025}{729} \\
 &= 5.646
 \end{aligned}$$

2. We are 95% confident that the ratio of the variances in the wait times at the two clinics is between 1.416 and 5.646.
3. Because 1 is outside the confidence interval, it suggests that the ratio of the variances $\frac{\sigma_1^2}{\sigma_2^2}$ is not 1. If the ratio of the variances cannot equal 1, then the variances cannot be equal. So there is a difference in the variances of the wait times at the two clinics.

NOTES

1. When calculating the limits for the confidence interval, keep all of the decimals in the F -scores and other values throughout the calculation. This will ensure that there is no round-off error in the answer. Use Excel to do the calculations of the limits, clicking on the cells containing the F -scores and any other values.
2. When writing down the interpretation of the confidence interval, make sure to include the confidence level and the actual ratio of population variances captured by the confidence interval (i.e. be specific to the context of the question).

TRY IT

A restaurateur knows that the mean revenue at her restaurant is higher on weekends than on weekdays. Now, she wants to investigate the variability in the revenue between weekends and weekdays. In a sample of 13 weekday orders, the variance in the prices of the orders was 63. In a sample of 11 weekend orders, the variance in the prices of the orders was 181.25.

1. Construct a 99% confidence interval for the ratio of the variances in the prices between weekdays and weekends.
2. Interpret the confidence interval found in part 1.
3. Is there evidence to suggest that the variances in the prices is the same on weekdays and weekends? Explain.

Click to see Solution

Let weekday prices be population 1 and weekend prices be population 2.

1.

Function	f.inv
Field 1	0.005
Field 2	12
Field 2	10
Answer	0.1966...

Function	f.inv.rt
Field 1	0.005
Field 2	12
Field 3	10
Answer	5.661...

$$\begin{aligned}
 \text{Lower Limit} &= \frac{1}{F_R} \times \frac{s_1^2}{s_2^2} \\
 &= \frac{1}{5.661\dots} \times \frac{63}{181.25} \\
 &= 0.0614
 \end{aligned}$$

$$\begin{aligned}
 \text{Upper Limit} &= \frac{1}{F_L} \times \frac{s_1^2}{s_2^2} \\
 &= \frac{1}{0.1966\dots} \times \frac{63}{181.25} \\
 &= 1.7676
 \end{aligned}$$

2. We are 99\% confident that the ratio of the variances in the prices on weekdays and weekends is between 0.0614 and 1.7676.
3. Because 1 is inside the confidence interval, it suggests that the ratio of the variances $\frac{\sigma_1^2}{\sigma_2^2}$ is equal to 1. If the ratio of the variances equals 1, then the variances must be equal. So the variances in the prices on weekdays and weekends are the same.

Conducting a Hypothesis Test for Two Population Variances

Follow these steps to perform a hypothesis test on two population variances:

1. Write down the null hypothesis that there is no difference in the population variances:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

The null hypothesis is always the claim that the two population variances are equal.

2. Write down the alternative hypotheses in terms of the difference in the population variances. The alternative hypothesis will be **one** of the following:

$$H_a : \sigma_1^2 < \sigma_2^2$$

$$H_a : \sigma_1^2 > \sigma_2^2$$

$$H_a : \sigma_1^2 \neq \sigma_2^2$$

3. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
4. Collect the sample information for the test and identify the significance level α .
5. Use the F -distribution to find the p – value (the area in the corresponding tail) for the test. The F -score and degrees of freedom are

$$F = \frac{s_1^2}{s_2^2}$$

$$df_1 = n_1 - 1$$

$$df_2 = n_2 - 1$$

6. Compare the p – value to the significance level and state the outcome of the test.
 - If p – value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If p – value $> \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
7. Write down a concluding sentence specific to the context of the question.

EXAMPLE

Two college instructors are interested in whether or not there is any variation in the way they grade math exams. They each grade the same set of 30 exams. The first instructor's grades have a variance of 52.3. The second instructor's grades have a variance of 89.9. At the 5% significance level, test the claim that the first instructor's variance is smaller.

Solution

Let the first instructor's grades be population 1 and the second instructor's grades be population 2. From the question, we have the following information:

Instructor 1	Instructor 2
$n_1 = 30$	$n_2 = 30$
$s_1^2 = 52.3$	$s_2^2 = 89.9$

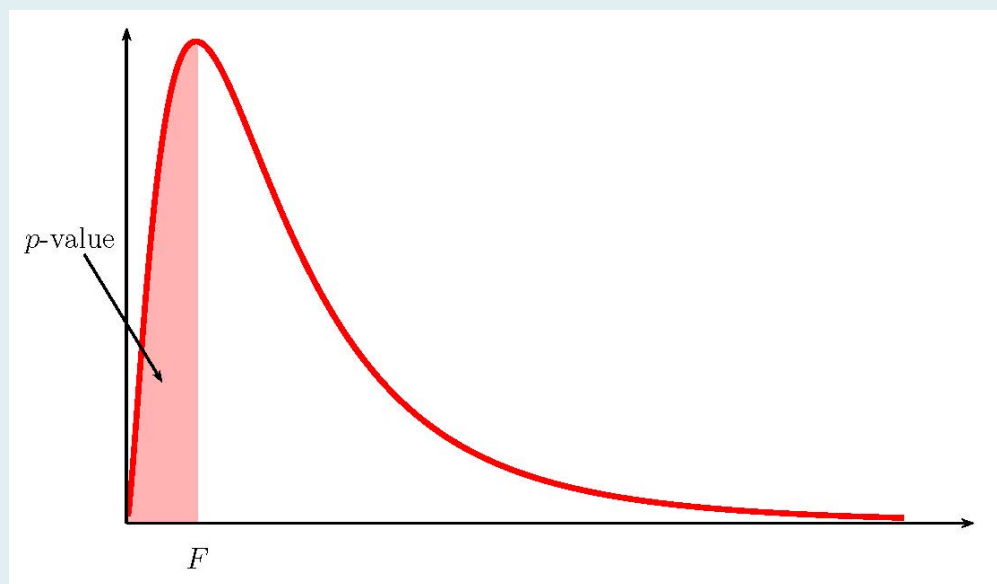
Hypotheses:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 < \sigma_2^2$$

p – value:

Because the alternative hypothesis is a $<$, the p – value is the area in the left tail of the F -distribution.



To use the **f.dist** function, we need to calculate the F -score and the degrees of freedom:

$$\begin{aligned}
 F &= \frac{s_1^2}{s_2^2} \\
 &= \frac{52.3}{89.9} \\
 &= 0.58175\dots
 \end{aligned}$$

$$\begin{aligned}
 df_1 &= n_1 - 1 \\
 &= 30 - 1 \\
 &= 29
 \end{aligned}$$

$$\begin{aligned}
 df_2 &= n_2 - 1 \\
 &= 30 - 1 \\
 &= 29
 \end{aligned}$$

Function	f.dist
Field 1	0.58175...
Field 2	29
Field 3	29
Field 4	true
Answer	0.0753

So the p – value = 0.0753.

Conclusion:

Because p – value = 0.0753 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level, there is not enough evidence to suggest that the first instructor's variance is smaller.

NOTES

1. The null hypothesis $\sigma_1^2 = \sigma_2^2$ is the claim that the variances for the two instructors are equal.
2. The alternative hypothesis $\sigma_1^2 < \sigma_2^2$ is the claim that the variance for the first instructor's grades is less than the variance for the second instructor's grades.
3. The p – value is the area in the left tail of the F -distribution, to the left of $F = 0.5817...$. In the calculation of the p – value:
 - The function is **f.dist** because we are finding the area in the left tail of an F -distribution.
 - Field 1 is the value of F .
 - Field 2 is the value of df_1 .
 - Field 3 is the value of df_2 .
 - Field 4 is true.
4. The p – value of 0.0753 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the variances for the two instructors are most likely

equal.

EXAMPLE

A local choral society divides the male singers into tenors and basses. The choral society director wants to know if the variance in the heights of the two groups of singers is the same or different. The director takes a sample from each group and records their height in inches. In a sample of 22 tenors, the sample variance is 3.89. In a sample of 27 basses, the sample variance is 2.72. At the 5% significance level, is there a difference in the heights of the two groups of singers?

Solution

Let the tenors be population 1, and the basses be population 2. From the question, we have the following information:

Tenors	Basses
$n_1 = 22$	$n_2 = 27$
$s_1^2 = 3.89$	$s^2 = 2.72$

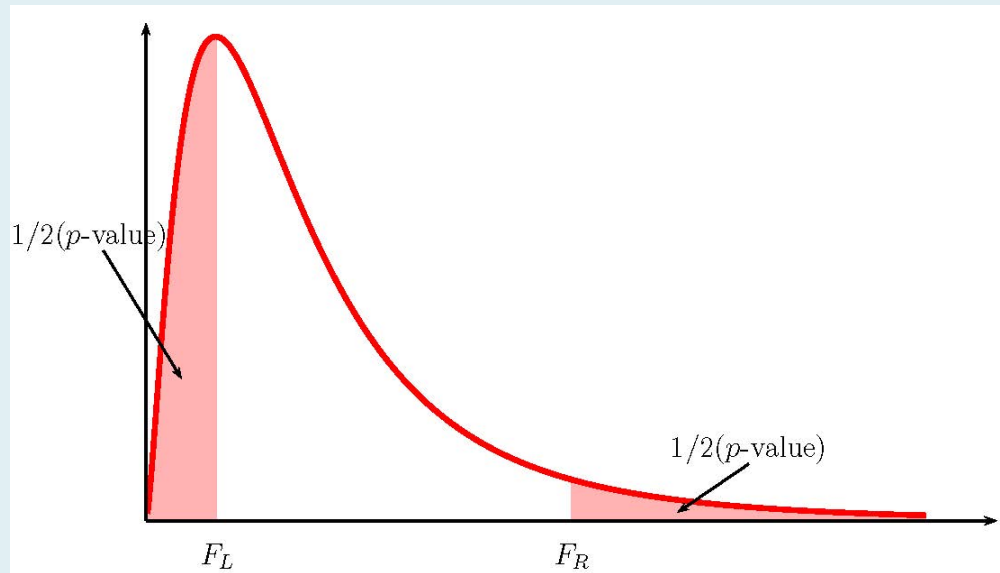
Hypotheses:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 \neq \sigma_2^2$$

p – value:

Because the alternative hypothesis is \neq , the p – value is the sum of the areas in the tails of the F -distribution.



We need to calculate out the F -score and the degrees of freedom:

$$\begin{aligned}
 F &= \frac{s_1^2}{s_2^2} \\
 &= \frac{3.89}{2.72} \\
 &= 1.430\dots
 \end{aligned}$$

$$\begin{aligned}
 df_1 &= n_1 - 1 \\
 &= 22 - 1 \\
 &= 21
 \end{aligned}$$

$$\begin{aligned}
 df_2 &= n_2 - 1 \\
 &= 27 - 1 \\
 &= 26
 \end{aligned}$$

Because this is a two-tailed test, we need to know which tail (left or right) we have the F -score belongs to so that we can use the correct Excel function. If $F > 1$, the F -score corresponds to the right tail. If the $F < 1$, the F -score corresponds to the left tail. In this case, $F = 1.430\dots > 1$, so the F -score corresponds to the right tail. We need to use **f.dist.rt** to find the area in the right tail.

Function	f.dist.rt
Field 1	1.430....
Field 2	21
Field 3	26
Answer	0.1919

So the area in the right tail is 0.1919, which means that $\frac{1}{2}p - \text{value} = 0.1919$. This is also the area in the left tail, so

$$p - \text{value} = 0.1919 + 0.1919 = 0.3838$$

Conclusion:

Because $p - \text{value} = 0.3838 > 0.05 = \alpha$, we do not reject the null hypothesis. At the 5% significance level, there is not enough evidence to suggest that there is a difference in the variation in the heights of the two groups of singers.

NOTES

1. The null hypothesis $\sigma_1^2 = \sigma_2^2$ is the claim that the variances of the heights for the two groups of singers are equal.
2. The alternative hypothesis $\sigma_1^2 \neq \sigma_2^2$ is the claim that the variances of the heights for the two groups of singers are not equal
3. In a two-tailed hypothesis test for two population variances, we will only have sample information relating to **one** of the two tails. We must determine which of the tails the sample information belongs to, and then calculate out the area in that tail. The area in each tail represents exactly half of the $p - \text{value}$, so the $p - \text{value}$ is the sum of the areas in the two tails.
 - If $F < 1$, the sample information belongs to the **left tail**.
 - We use **f.dist** to find the area in the left tail. The area in the right tail equals the area in the left tail, so we can find the $p - \text{value}$ by adding the output from this function to itself.
 - If $F > 1$, the sample information belongs to the **right tail**.

- We use **f.dist.rt** to find the area in the right tail. The area in the left tail equals the area in the right tail, so we can find the **p — value** by adding the output from this function to itself.
4. The **p — value** of **0.3838** is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the variances in the heights of the two groups of singers are the same.

NOTES

1. When two populations have equal variances, the values of s_1^2 and s_2^2 are close in value. So, the value of $\frac{s_1^2}{s_2^2}$ is close to 1. This will result in a large **p — value** in the hypothesis test, and the evidence favours the null hypothesis.
2. When two populations have unequal variances, then the values of s_1^2 and s_2^2 are not close in value. So, the value of $\frac{s_1^2}{s_2^2}$ will either be larger than 1 or smaller than 1 (depending on which sample variance is smaller and which is larger). This will result in a small **p — value** in the hypothesis test, and the evidence favours the alternative hypothesis.

TRY IT

A researcher knows that the mean annual repair costs for a car increases with the age of the car. But what about the variance in the repair costs? The researcher wants to know if the variance in the annual repair costs for cars increases as cars get older. In a sample of 28 older (5 years or more) cars, the standard deviation of the annual repair costs was \$150. In a sample of 25 newer (3 years or less) cars, the standard deviation of the annual repair costs was \$90. At the 5% significance level, test if the variance in the annual repair costs is larger for older cars than for newer cars.

Click to see Solution

Let the older cars be population 1, and the newer cars be population 2.

Hypotheses:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 > \sigma_2^2$$

p – value:

From the question, we have $n_1 = 28$, $s_1^2 = 22,500$, $n_2 = 25$, $s_2^2 = 8,100$, and $\alpha = 0.05$.

Because the alternative hypothesis is a $>$, the p – value is the area in the right tail of the F -distribution.

To use the **f.dist.rt** function, we need to calculate out the F -score and the degrees of freedom:

$$\begin{aligned}
 F &= \frac{s_1^2}{s_2^2} \\
 &= \frac{22,500}{8,100} \\
 &= 2.777 \dots
 \end{aligned}$$

$$\begin{aligned}
 df_1 &= n_1 - 1 \\
 &= 28 - 1 \\
 &= 27
 \end{aligned}$$

$$\begin{aligned}
 df_2 &= n_2 - 1 \\
 &= 25 - 1 \\
 &= 24
 \end{aligned}$$

Function	f.dist.rt
Field 1	2.777...
Field 2	27
Field 3	24
Answer	0.0068

So the p – value = 0.0068.

Conclusion:

Because p – value = 0.0068 < 0.05 = α , we reject the null hypothesis. At the 5\% significance level there is enough evidence to suggest that the variance in the annual repair costs is larger for older cars than for newer cars.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=264#oembed-1>

Video: “Hypothesis Tests for Equality of Two Variances” by jbstatistics [11:40] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. As part of her capstone project, a business student is studying the starting salaries of accounting graduates and finance graduates. In a sample of 35 accounting graduates, the starting salaries had a variance of 137.2. In a sample of 30 finance graduates, the starting salaries had a variance of 60.4.
 - a. Construct a 95% confidence interval for the ratio of the variances of the starting salaries of accounting and finance graduates.
 - b. Interpret the confidence interval found in part (a).
 - c. Is there evidence to suggest that the variance in the starting salaries of accounting graduates is higher than that of finance graduates? Explain.

Click to see Answer

- a. Lower Limit = 1.101, Upper Limit = 4.590
 - b. There is a 95% probability that the ratio of the variances in the starting salaries of accounting and finance graduates is between 1.101 and 4.590.
 - c. Yes. Because the lower limit is greater than 1, it suggests that the ratio of the two variances is greater than 1. This can only happen if the variance for population 1 (accounting) is greater than the variance for population 2 (finance).
2. Two co-workers commute from the same building. They are interested in whether or not there is any variation in the time it takes them to drive to work. They each record their times for 20 commutes. The first worker's times have a variance of 12.1. The second worker's times have a variance of 16.9. At the 5% significance level, test if the variation in the first worker's commute time is smaller than the second worker's.

Click to see Answer

- Hypotheses: $H_0 : \sigma_1^2 = \sigma_2^2$
 $H_a : \sigma_1^2 < \sigma_2^2$
- p – value = 0.2367
- Conclusion: At the 5% significance level, there is not enough evidence to conclude that the variation in the first worker's commute time is smaller than the variation in the

second worker's.

3. Two students are interested in whether or not there is variation in their test scores for math class. So far, they each have taken a total of 15 total math tests. The first student's grades have a standard deviation of 38.1. The second student's grades have a standard deviation of 22.5. At the 5% significance level, determine if the variation in the second student's scores are lower than the first student's.

Click to see Answer

- Hypotheses: $H_0 : \sigma_1^2 = \sigma_2^2$
 $H_a : \sigma_1^2 > \sigma_2^2$
- p – value = 0.0291
- Conclusion: At the 5% significance level, there is enough evidence to conclude that the variation in the second student's scores is lower than the variation in the first student's.

4. Two cyclists are comparing the variances of their overall paces going uphill. In a sample of 35 hill climbs, the first cyclist's speeds have a variance of 23.8. In a sample of 40 hill climbs, the second cyclist's speeds have a variance of 37.6. At the 5% significance level, is there a difference in the variance in the cyclists' speeds?

Click to see Answer

- Hypotheses: $H_0 : \sigma_1^2 = \sigma_2^2$
 $H_a : \sigma_1^2 \neq \sigma_2^2$
- p – value = 0.1775
- Conclusion: At the 5% significance level, there is not enough evidence to conclude that there is a difference in the variance in the cyclists' speeds.

5. Students Linda and Tuan are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula *A*, and Tuan feeds his rats Formula *B*. At the end of a specified time period, each rat is weighed again, and the net gain in grams is recorded.

Linda's Rats	Tuan's Rats
43.5	47.0
39.4	40.5
41.3	38.9
46.0	46.3
38.2	44.2

- Construct a 98% confidence interval for the ratio of the variances in the net weight gain for Linda's and Tuan's rats.
- Interpret the confidence interval found in part (a).
- Is there evidence to suggest that the variance in the net weight gain for Linda and Tuan's rats is the same? Explain.

Click to see Answer

- Lower Limit = 0.049, Upper Limit = 12.433
 - There is a 98% probability that the ratio of the variances in the net weight gain for Linda's and Tuan's rats is between 0.049 and 12.433.
 - Yes. Because 1 is inside the confidence interval, it suggests that the ratio of the two variances equals 1. If the ratio of the variances equals 1, then the variances must be equal.
6. A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most because they commute the farthest to work. Suppose that the group randomly surveyed 16 individuals and asked them their daily one-way commuting distance, in kilometres. The results are shown in the table below.

Working-Class	Professional (Middle Incomes)
17.8	16.5
26.7	17.4
49.4	22.0
9.4	7.4
65.4	9.4
47.1	2.1
19.5	6.4
51.2	13.9

Determine whether or not the variance in kilometres driven to work is statistically the same among the working class and professional (middle-income) groups. Use a 5% significance level.

Click to see Answer

- Hypotheses: $H_0 : \sigma_1^2 = \sigma_2^2$
 $H_a : \sigma_1^2 \neq \sigma_2^2$
- p – value = 0.0160
- Conclusion: At the 5% significance level, there is enough evidence to conclude that there is a difference in the variance in the kilometres driven to work for the two groups.

7. A researcher wants to study the variation in the amount of money, in dollars, that shoppers spend on Saturdays and Sundays at the mall. In a sample of 12 Saturday shoppers, the standard deviation for the amount of money shoppers spent was \$43.2. In a sample of 16 Sunday shoppers, the standard deviation for the amount of money shoppers spent was \$38.3.
- Construct a 93% confidence interval for the ratio of the variances for the amount of money spent by shoppers on Saturdays and Sundays at the mall.
 - Interpret the confidence interval found in part (a).
 - Is there evidence to suggest that variance in the amount of money spent on Saturdays and Sundays at the mall is different? Explain.

Click to see Answer

- Lower Limit = 0.461, Upper Limit = 3.848
- There is a 93% probability that the ratio of the variance in the amount of money spent by shoppers on Saturdays and Sundays at the mall is between 0.461 and 3.848.

- c. No. Because 1 is inside the confidence interval, it suggests that the ratio of the two variances equals 1. If the ratio of the variances equals 1, then the variances must be equal.

8. A researcher is studying the incomes of people who live on the East Coast or West Coast. The table shows the results of the study. Income is shown in thousands of dollars. Assume that both distributions are normal.

East	West
38	71
47	126
30	42
82	51
75	44
52	90
115	88
67	

At the 5% significance level, determine if the variation in income for people who live on the East Coast is lower than for people who live on the West Coast.

Click to see Answer

- Hypotheses: $H_0 : \sigma_1^2 = \sigma_2^2$
 $H_a : \sigma_1^2 < \sigma_2^2$
- p – value = 0.1632
- Conclusion: At the 5% significance level, there is not enough evidence to conclude that the variation in income for people who live on the East Coast is lower than for people who live on the West Coast.

9. A financial planner is considering investing her client's money into one of two possible stock options, A and B . The client is risk adverse and wants to invest her money into a stock with the least amount of variability in the monthly percentage return of the stock. In a random sample of 20 months of returns of stock A , the standard deviation is 31.7%. In a random

sample of 30 months of returns of stock B , the standard deviation is 19.5\%.

- a. At the 1\% significance level, is the variation in the monthly returns of stock A greater than the variation in the monthly returns of stock B ?
- b. Which stock should the financial planner invest the client's money? Why?

Click to see Answer

- Hypotheses: $H_0 : \sigma_1^2 = \sigma_2^2$
 $H_a : \sigma_1^2 > \sigma_2^2$
- p – value = 0.0090
- Conclusion: At the 1\% significance level, there is enough evidence to conclude that the variation in the monthly returns of stock A is greater than the variation in the monthly returns of stock B .

- b. Stock B because it has the smaller variation in the monthly returns.

10. One of the measures of the quality of a production process is the variance in the process. A smaller variance means that there is more consistency in the product. A larger variance indicates that there is less consistency in the product. A company that produces 500 gram bags of coffee uses two different machines in the bagging process. The company takes a sample of bags produced by each machine and records the weight, in grams, of each bag. The results are given in the table below.

Machine 1	Machine 2
495.5	501.3
502.7	499.9
495.9	500.7
498.6	498.2
501.6	500.4
502.5	498.7
499.2	499.2
496.7	496.3
498.7	498.1
499.0	500.1
499.3	499.8
498.8	499.3
499.1	500.2
500.8	500.1
499.3	497.5
497.2	499.6
498.1	499.6
499.4	497.7
499.2	499.6
498.2	498.8
497.7	499.7
498.8	
500.3	

At the 1\% significance level, is there a difference in the variance of the weights of the bags produced by each machine?

Click to see Answer

- Hypotheses: $H_0 : \sigma_1^2 = \sigma_2^2$
 $H_a : \sigma_1^2 \neq \sigma_2^2$
- $p - \text{value} = 0.0652$

- **Conclusion:** At the 1\% significance level, there is not enough evidence to conclude that there is a difference in the variance of the weights of the bags produced by the two machines.

“11.3 Statistical Inference for Two Population Variances” and “11.5 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

11.3 ONE-WAY ANOVA AND HYPOTHESIS TESTS FOR THREE OR MORE POPULATION MEANS

LEARNING OBJECTIVES

- Conduct and interpret hypothesis tests for three or more population means using one-way ANOVA.

The purpose of a one-way ANOVA (analysis of variance) test is to determine the existence of a statistically significant difference among the means of three or more populations. The test actually uses variances to help determine if the population means are equal or not.

Throughout this section, we will use subscripts to identify the values for the means, sample sizes, and standard deviations for the populations.

Symbol for:	Population k
Population Mean	μ_k
Population Standard Deviation	σ_k
Sample Size	n_k
Sample Mean	\bar{x}_k
Sample Standard Deviation	s_k

k is the number of populations under study, n is the total number of observations in all of the samples combined, and $\bar{\bar{x}}$ is the mean of the sample means.

$$n = n_1 + n_2 + \cdots + n_k$$

$$\bar{\bar{x}} = \frac{n_1 \times \bar{x}_1 + n_2 \times \bar{x}_2 + \cdots + n_k \times \bar{x}_k}{n}$$

One-Way ANOVA

A predictor variable is called a **factor** or **independent variable**. For example, age, temperature, and gender are factors. The groups or samples are often referred to as **treatments**. This terminology comes from the use of ANOVA procedures in medical and psychological research to determine if there is a difference in the effects of different treatments.

EXAMPLE

A local college wants to compare the mean GPA for players on four of its sports teams: basketball, baseball, hockey, and lacrosse. A random sample of players was taken from each team, and their GPAs were recorded in the table below.

Basketball	Baseball	Hockey	Lacrosse
3.6	2.1	4.0	2.0
2.9	2.6	2.0	3.6
2.5	3.9	2.6	3.9
3.3	3.1	3.2	2.7
3.8	3.4	3.2	2.5

In this example, the factor is the sports team.

	Basketball	Baseball	Hockey	Lacrosse
	Population 1	Population 2	Population 3	Population 4
Sample Size (n_i)	5	5	5	5
Sample Mean (\bar{x}_i)	3.22	3.02	3	2.94

$$k = 4$$

$$\begin{aligned} n &= n_1 + n_2 + n_3 + n_4 \\ &= 5 + 5 + 5 + 5 \\ &= 20 \end{aligned}$$

$$\begin{aligned} \bar{\bar{x}} &= \frac{n_1 \times \bar{x}_1 + n_2 \times \bar{x}_2 + n_3 \times \bar{x}_3 + n_4 \times \bar{x}_4}{n} \\ &= \frac{5 \times 3.22 + 5 \times 3.02 + 5 \times 3 + 5 \times 2.94}{20} \\ &= 3.045 \end{aligned}$$

The following assumptions are required to use a one-way ANOVA test:

1. Each population from which a sample is taken is normally distributed.
2. All samples are randomly selected and independently taken from the populations.
3. The populations are assumed to have **equal variances**.
4. The population data is numerical (interval or ratio level).

The logic behind one-way ANOVA is to compare population means based on two independent estimates of the (assumed) equal variance σ^2 between the populations:

- One estimate of the equal variance σ^2 is based on the variability among the sample means themselves (called the between-groups estimate of population variance).
- One estimate of the equal variance σ^2 is based on the variability of the data within each sample (called the within-groups estimate of population variance).

The one-way ANOVA procedure compares these two estimates of the population variance σ^2 to determine if the population means are equal or if there is a difference in the population means.

Because ANOVA involves the comparison of two estimates of variance, an F -distribution is used to conduct the ANOVA test. The test statistic is an F -score that is the ratio of the two estimates of population variance:

$$F = \frac{\text{variance between groups}}{\text{variance within groups}}$$

The degrees of freedom for the F -distribution are $df_1 = k - 1$ and $df_2 = n - k$ where k is the number of populations and n is the total number of observations in all of the samples combined.

The **variance between groups** estimate of the population variance is called the **mean square due to treatment**, MST . The MST is the estimate of the population variance determined by the variance of the sample means from the overall sample mean $\bar{\bar{x}}$. When the population means are equal, MST provides an unbiased estimate of the population variance. When the population means are not equal, MST provides an overestimate of the population variance.

$$SST = n_1 \times (\bar{x}_1 - \bar{\bar{x}})^2 + n_2 \times (\bar{x}_2 - \bar{\bar{x}})^2 + \cdots + n_k \times (\bar{x}_k - \bar{\bar{x}})^2$$

$$MST = \frac{SST}{k - 1}$$

The **variance within groups** estimate of the population variance is called the **mean square due to error**, MSE . The MSE is the pooled estimate of the population variance using the sample variances as estimates for the population variance. The MSE always provides an unbiased estimate of the population variance because it is not affected by whether or not the population means are equal.

$$SSE = (n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2 + \cdots + (n_k - 1) \times s_k^2$$

$$MSE = \frac{SSE}{n - k}$$

The one-way ANOVA test depends on the fact that the variance between groups MST is influenced by differences between the population means, which results in MST being either an unbiased or overestimate of the population variance. Because the variance within groups MSE compares values of each group to its own group mean, MSE is not affected by differences between the population means and is always an unbiased estimate of the population variance.

The null hypothesis in a one-way ANOVA test is that the population means are all equal, and the alternative hypothesis is that there is a difference in the population means. The F -score for the one-

way ANOVA test is $F = \frac{MST}{MSE}$ with $df_1 = k - 1$ and $df_2 = n - k$. The p - value for the test is the area in the right tail of the F -distribution, to the right of the F -score.

When the variance between groups MST and variance within groups MSE are close in value, the F -score is close to 1 and results in a large p - value. In this case, the conclusion is that the population means are equal.

When the variance between groups MST is significantly larger than the variability within groups MSE , the F -score is large and results in a small p - value. In this case, the conclusion is that there is a difference in the population means.

Conducting a Hypothesis Test for Three or More Population Means

Follow these steps to perform a hypothesis test on three or more population means:

1. Verify that the one-way ANOVA assumptions are met.
2. Write down the null hypothesis that there is no difference in the population means:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

The null hypothesis is always the claim that the population means are equal.

3. Write down the alternative hypotheses that there is some difference in the population means:

$$H_a : \text{at least one population mean is different from the others}$$

4. Collect the sample information for the test and identify the significance level α .
5. The p - value is the area in the right tail of the F -distribution. The F -score and degrees of freedom are

$$F = \frac{MST}{MSE}$$

$$df_1 = k - 1$$

$$df_2 = n - k$$

6. Compare the p – value to the significance level and state the outcome of the test.
 - If p – value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If p – value $> \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
7. Write down a concluding sentence specific to the context of the question.

EXAMPLE

A local college wants to compare the mean GPA for players on four of its sports teams: basketball, baseball, hockey, and lacrosse. A random sample of players was taken from each team, and their GPA was recorded in the table below.

Basketball	Baseball	Hockey	Lacrosse
3.6	2.1	4.0	2.0
2.9	2.6	2.0	3.6
2.5	3.9	2.6	3.9
3.3	3.1	3.2	2.7
3.8	3.4	3.2	2.5

Assume the populations are normally distributed and have equal variances. At the 5% significance level, is there a difference in the average GPA between the sports team?

Solution

Let basketball be population 1, let baseball be population 2, let hockey be population 3, and let lacrosse be population 4. From the question, we have the following information:

Basketball	Baseball	Hockey	Lacrosse
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$	$n_4 = 5$
$\bar{x}_1 = 3.22$	$\bar{x}_2 = 3.02$	$\bar{x}_3 = 3$	$\bar{x}_4 = 2.94$
$s_1^2 = 0.277$	$s_2^2 = 0.487$	$s_3^2 = 0.56$	$s_4^2 = 0.613$

Previously, we found $k = 4$, $n = 20$, and $\bar{\bar{x}} = 3.045$.

Hypotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

H_a : at least one population mean is different from the others

p – value:

To calculate out the F -score, we need to find MST and MSE .

$$\begin{aligned}
 SST &= n_1 \times (\bar{x}_1 - \bar{\bar{x}})^2 + n_2 \times (\bar{x}_2 - \bar{\bar{x}})^2 + n_3 \times (\bar{x}_3 - \bar{\bar{x}})^2 + n_4 \times (\bar{x}_4 - \bar{\bar{x}})^2 \\
 &= 5 \times (3.22 - 3.045)^2 + 5 \times (3.02 - 3.045)^2 + 5 \times (3 - 3.045)^2 \\
 &\quad + 5 \times (2.94 - 3.045)^2 \\
 &= 0.2215
 \end{aligned}$$

$$\begin{aligned}
 MST &= \frac{SST}{k - 1} \\
 &= \frac{0.2215}{4 - 1} \\
 &= 0.0738\dots
 \end{aligned}$$

$$\begin{aligned}
 SSE &= (n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2 + (n_3 - 1) \times s_3^2 + (n_4 - 1) \times s_4^2 \\
 &= (5 - 1) \times 0.277 + (5 - 1) \times 0.487 + (5 - 1) \times 0.56 + (5 - 1) \times 0.623 \\
 &= 7.788
 \end{aligned}$$

$$\begin{aligned}
 MSE &= \frac{SSE}{n - k} \\
 &= \frac{7.788}{20 - 4} \\
 &= 0.48675
 \end{aligned}$$

The p – value is the area in the right tail of the F -distribution. To use the **f.dist.rt** function, we need to calculate out the F -score and the degrees of freedom:

$$\begin{aligned} F &= \frac{MST}{MSE} \\ &= \frac{0.0738\dots}{0.48675} \\ &= 0.15168\dots \end{aligned}$$

$$\begin{aligned} df_1 &= k - 1 \\ &= 4 - 1 \\ &= 3 \end{aligned}$$

$$\begin{aligned} df_2 &= n - k \\ &= 20 - 4 \\ &= 16 \end{aligned}$$

Function	f.dist.rt
Field 1	0.15168...
Field 2	3
Field 3	16
Answer	0.9271

So the p – value = 0.9271.

Conclusion:

Because p – value = 0.9271 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level, there is enough evidence to suggest that the mean GPA for the sports teams are the same.

NOTES

1. The null hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4$ is the claim that the mean GPA for the sports teams are all equal.

2. The alternative hypothesis is the claim that at least one of the population means is not equal to the others. The alternative hypothesis does not say that all of the population means are not equal, only that at least one of them is not equal to the others.
3. The **p – value** is the area in the right tail of the F -distribution, to the right of $F = 0.15168 \dots$. In the calculation of the **p – value**:
 - The function is **f.dist.rt** because we are finding the area in the right tail of an F -distribution.
 - Field 1 is the value of F .
 - Field 2 is the value of df_1 .
 - Field 3 is the value of df_2 .
4. The **p – value** of **0.9271** is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the population means are all equal.

ANOVA Summary Tables

The calculation of the MST , MSE , and the F -score for a one-way ANOVA test can be time-consuming, even with the help of software like Excel. However, Excel has a built-in one-way ANOVA summary table that not only generates the averages, variances, MST , and MSE , but also calculates the required F -score and **p – value** for the test.

USING EXCEL TO CREATE A ONE-WAY ANOVA SUMMARY TABLE

In order to create a one-way ANOVA summary table, we need to use the Analysis ToolPak. Follow these instructions to add the Analysis ToolPak.

1. Enter the data into an Excel worksheet.
2. Go to the **Data** tab and click on **Data Analysis**. If you do not see **Data Analysis** in the **Data** tab, you will need to install the Analysis ToolPak.
3. In the **Data Analysis** window, select **Anova: Single Factor**. Click **OK**.
4. In the **Input** range, enter the cell range for the data.
5. In the **Grouped By** box, select rows if your data is entered as rows (the default is columns).
6. Click on **Labels in first row** if you included the column headings in the input range.
7. In the **Alpha** box, enter the significance level for the test.
8. From the **Output Options**, select the location where you want the output to appear.
9. Click **OK**.

This website provides additional information on using Excel to create a one-way ANOVA summary table.

NOTE

Because we are using the p — value approach to hypothesis testing, it is not crucial that we enter the actual significance level we are using for the test. The p — value (the area in the right tail of the F -distribution) is not affected by significance level. For the critical-value approach to hypothesis testing, we must enter the correct significance level for the test because the critical value does depend on the significance level.

EXAMPLE

A local college wants to compare the mean GPA for players on four of its sports teams: basketball, baseball, hockey, and lacrosse. A random sample of players was taken from each team, and their GPA was recorded in the table below.

Basketball	Baseball	Hockey	Lacrosse
3.6	2.1	4.0	2.0
2.9	2.6	2.0	3.6
2.5	3.9	2.6	3.9
3.3	3.1	3.2	2.7
3.8	3.4	3.2	2.5

Assume the populations are normally distributed and have equal variances. At the 5% significance level, is there a difference in the average GPA between the sports team?

Solution

Let basketball be population 1, let baseball be population 2, let hockey be population 3, and let lacrosse be population 4.

Hypotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

H_a : at least one population mean is different from the others

p – value:

The ANOVA summary table generated by Excel is shown below:

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Basketball	5	16.1	3.22	0.277		
Baseball	5	15.1	3.02	0.487		
Hockey	5	15	3	0.56		
Lacrosse	5	14.7	2.94	0.623		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.2215	3	0.073833	0.151686	0.927083	3.238872
Within Groups	7.788	16	0.48675			
Total	8.0095	19				

The p – value for the test is in the **P-value column** of the **between groups row**. So the p – value = 0.9271.

Conclusion:

Because p – value = 0.9271 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level, there is enough evidence to suggest that the mean GPA for the sports teams are the same.

NOTES

1. In the top part of the ANOVA summary table (under the Summary heading), we have the averages and variances for each of the groups (basketball, baseball, hockey, and lacrosse).
2. In the bottom part of the ANOVA summary table (under the ANOVA heading), we have

- The value of SST (in the SS column of the **between groups** row).
- The value of MST (in the MS column of the **between groups** row).
- The value of SSE (in the SS column of the **within groups** row).
- The value of MSE (in the MS column of the **within groups** row).
- The value of the F -score (in the F column of the **between groups** row).
- The p — value (in the p — value column of the **between groups** row).

EXAMPLE

A fourth-grade class is studying the environment. One of the assignments is to grow bean plants in different soils. Tommy chose to grow his bean plants in soil found outside his classroom mixed with dryer lint. Tara chose to grow her bean plants in potting soil bought at the local nursery. Nick chose to grow his bean plants in soil from his mother's garden. No chemicals were used on the plants, only water. They were grown inside the classroom next to a large window. Each child grew five plants. At the end of the growing period, each plant was measured, producing the data (in inches) in the table below.

Tommy's Plants	Tara's Plants	Nick's Plants
24	25	23
21	31	27
23	23	22
30	20	30
23	28	20

Assume the heights of the plants are normally distribution and have equal variance. At the 5\%

significance level, does it appear that the three media in which the bean plants were grown produced the same mean height?

Solution

Let Tommy's plants be population 1, let Tara's plants be population 2, and let Nick's plants be population 3.

Hypotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

H_a : at least one population mean is different from the others

p – value:

The ANOVA summary table generated by Excel is shown below:

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Tommy's Plants	5	121	24.2	11.7		
Tara's Plants	5	127	25.4	18.3		
Nick's Plants	5	122	24.4	16.3		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	4.133333	2	2.066667	0.133909	0.875958	3.885294
Within Groups	185.2	12	15.43333			
Total	189.3333	14				

So the p – value = 0.8760.

Conclusion:

Because p – value = 0.8760 > 0.05 = α , we do not reject the null hypothesis. At the 5%

significance level, there is enough evidence to suggest that the mean heights of the plants grown in three media are the same.

NOTES

1. The null hypothesis $\mu_1 = \mu_2 = \mu_3$ is the claim that the mean heights of the plants grown in the three different media are all equal.
2. The alternative hypothesis is the claim that at least one of the population means is not equal to the others. The alternative hypothesis does not say that all of the population means are not equal, only that at least one of them is not equal to the others.
3. The p — value of 0.8760 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the population means are all equal.

TRY IT

A statistics professor wants to study the average GPA of students in four different programs: marketing, management, accounting, and human resources. The professor took a random sample of GPAs of students in those programs at the end of the past semester. The data is recorded in the table below.

Marketing	Management	Accounting	Human Resources
2.17	2.63	3.21	3.27
1.85	1.77	3.78	3.45
2.83	3.25	4.00	2.85
1.69	1.86	2.95	2.26
3.33	2.21	2.65	3.18

Assume the GPAs of the students are normally distributed and have equal variance. At the 5\% significance level, is there a difference in the average GPA of the students in the different programs?

Click to see Solution

Let marketing be population 1, let management be population 2, let accounting be population 3, and let human resources be population 4.

Hypotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

H_a : at least one population mean is different from the others

p – value:

The ANOVA summary table generated by Excel is shown below:

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Marketing	5	11.87	2.374	0.47648		
Management	5	11.72	2.344	0.37108		
Accounting	5	16.59	3.318	0.31797		
Human Resources	5	15.01	3.002	0.21947		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	3.459895	3	1.153298	3.330826	0.046214	3.238872
Within Groups	5.54	16	0.34625			
Total	8.999895	19				

So the p – value = 0.0462.

Conclusion:

Because p – value = 0.0462 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5\% significance level, there is enough evidence to suggest that there is a difference in the average GPA of the students in the different programs.

TRY IT

A manufacturing company runs three different production lines to produce one of its products. The company wants to know if the mean production rate is the same for the three lines. For each production line, a sample of eight-hour shifts was taken, and the number of items produced during each shift was recorded in the table below.

Line 1	Line 2	Line 3
35	21	31
35	36	34
36	22	24
39	38	21
37	28	27
36	34	29
31	35	33
38	39	20
33	40	24

Assume the numbers of items produced on each line during an eight-hour shift are normally distributed and have equal variance. At the 1% significance level, is there a difference in the average production rate for the three lines?

Click to see Solution

Let Line 1 be population 1, let Line 2 be population 2, and let Line 3 be population 3.

Hypotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

H_a : at least one population mean is different from the others

p – value:

The ANOVA summary table generated by Excel is shown below:

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Line 1	9	320	35.55556	6.027778		
Line 2	9	293	32.55556	51.52778		
Line 3	9	243	27	26		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	339.1852	2	169.5926	6.089096	0.007264	5.613591
Within Groups	668.4444	24	27.85185			
Total	1007.63	26				

So the p – value = 0.0073.

Conclusion:

Because p – value = 0.0073 < 0.01 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 1\% significance level, there is enough evidence to suggest that there is a difference in the mean production rate of the three lines.

Exercises

1. Three different traffic routes are tested for mean driving time. The entries in the table are the driving times, in minutes, on the three different routes. Assume the driving times are normally distribution and have equal variance.

Route 1	Route 2	Route 3
30	27	16
32	29	41
27	28	22
35	36	31

At the 5% significance level, test if the mean driving time for the three routes is the same.

Click to see Answer

- Hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3$
 $H_a : \text{at least one population mean is different from the others}$
- $p - \text{value} = 0.7728$
- Conclusion: At the 5% significance level, there is enough evidence to suggest that the mean driving time is the same for the three routes.

2. Suppose a group is interested in determining whether teenagers obtain their driver's licenses at approximately the same mean age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their driver's licenses. Assume the ages are normally distribution and have equal variance.

Northeast	South	West	Central	East
16.3	16.9	16.4	16.2	17.1
16.1	16.5	16.5	16.6	17.2
16.4	16.4	16.6	16.5	16.6
16.5	16.2	16.1	16.4	16.8

At the 5% significance level, determine if the mean age when teenagers get their driver's license is the same in the different regions of the country.

Click to see Answer

- Hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
 $H_a : \text{at least one population mean is different from the others}$
- $p - \text{value} = 0.0174$

- Conclusion: At the 5\% significance level, there is enough evidence to suggest that there is a difference in the mean age when teenagers get their driver's licenses in different regions of the country.

3. Groups of men from three different areas of the country are to be tested for mean weight. The entries in the table are the weights for the different groups. Assume the weights are normally distribution and have equal variance.

Group 1	Group 2	Group 3
216	202	170
198	213	165
240	284	182
187	228	197
176	210	201

At the 1\% significance level, test if the mean weight for men is the same for the three groups.

Click to see Answer

- Hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3$
 $H_a : \text{at least one population mean is different from the others}$
- $p - \text{value} = 0.0546$
- Conclusion: At the 1\% significance level, there is enough evidence to suggest that the mean weight for men is the same for the three groups.

4. Girls from four different soccer teams are tested for mean goals scored per game. The entries in the table are the goals per game for the different teams for a sample of games. Assume the goals scored per game are normally distribution and have equal variance.

Team 1	Team 2	Team 3	Team 4
1	2	0	3
2	3	1	4
0	2	1	4
3	4	0	3
2	4	0	2

At the 1\% significance level, test if the mean goals scored per game is the same for the four teams.

Click to see Answer

- Hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
 $H_a : \text{at least one population mean is different from the others}$
- $p - \text{value} = 0.0005$
- Conclusion: At the 1\% significance level, there is enough evidence to suggest that there is a difference in the mean goals scored per game for the four teams.

5. Five basketball teams took a random sample of players regarding how high each player can jump (in centimetres). Assume the heights are normally distribution and have equal variance.

Team 1	Team 2	Team 3	Team 4	Team 5
90	80	120	95	102.5
105	87.5	125	110	97.5
127.5	95	97.5	115	100

At the 5\% significance level, is there a difference in the mean jump heights among the teams?

Click to see Answer

- Hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
 $H_a : \text{at least one population mean is different from the others}$
- $p - \text{value} = 0.1614$
- Conclusion: At the 5\% significance level, there is enough evidence to suggest that there is no difference in the mean jump height among the teams.

6. A video game developer is testing a new game on three different groups. Each group represents a different target market for the game. The developer collects scores from a random sample from each group. The scores are recorded in the table below. Assume the scores are normally distribution and have equal variance.

Group A	Group B	Group C
101	151	101
108	149	109
98	160	198
107	112	186
111	126	160

At the 5\% significance level, test if the mean scores are the same for the different groups.

Click to see Answer

- Hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3$
 $H_a : \text{at least one population mean is different from the others}$
- $p - \text{value} = 0.0592$
- Conclusion: At the 5\% significance level, there is enough evidence to suggest that the mean scores are the same for the different groups.

7. Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula *A*, Tuan feeds his rats Formula *B*, and Javier feeds his rats Formula *C*. At the end of a specified time period, each rat is weighed again, and the net gain, in grams, is recorded. The results are shown in the table below. Assume the net weight gains are normally distribution and have equal variance.

Linda's Rats	Tuan's Rats	Javier's Rats
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

At the 5%, determine if the three formulas produce the same mean net weight gain.

Click to see Answer

- Hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3$
 $H_a : \text{at least one population mean is different from the others}$
- $p - \text{value} = 0.5305$
- Conclusion: At the 5% significance level, there is enough evidence to suggest that the mean net weight gain is the same for the three formulas.

8. A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most because they commute the farthest to work. Suppose that the group randomly surveyed 8 individuals in each of three different income groups (working-class, middle-income, and wealthy), and asked them their daily one-way commuting distance, in kilometres. Assume the distances are normally distribution and have equal variance.

Working-Class	Middle Income	Wealthy
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

At the 1% significance level, determine if there is a difference in the mean commuting distance for the three income groups.

Click to see Answer

- Hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3$
 $H_a : \text{at least one population mean is different from the others}$
- $p - \text{value} = 0.0014$
- Conclusion: At the 1% significance level, there is enough evidence to suggest that there is a difference in the mean commuting distance for the three income groups.

9. The following table lists the number of pages in four different types of magazines. Assume the number of pages are normally distribution and have equal variance.

Home Decorating	News	Health	Computer
172	87	82	104
286	94	153	136
163	123	87	98
205	106	103	207
197	101	96	146

At the 5\% significance level, test if the four magazine types have the same mean number of pages.

Click to see Answer

- Hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
 $H_a : \text{at least one population mean is different from the others}$
- $p - \text{value} = 0.0012$
- Conclusion: At the 5\% significance level, there is enough evidence to suggest that four magazine types do not have the same mean number of pages.

10. Are the means for the final exams the same for all statistics class delivery types? The table below shows the mean scores on final exams from several randomly selected classes that used the different delivery types. Assume the mean scores are normally distribution and have equal variance.

Online	Hybrid	Face-to-Face
72	83	80
84	73	78
77	84	84
80	81	81
81		86
		79
		82

At the 5% significance level, determine if the mean score on the final exam is the same for the different course delivery methods.

Click to see Answer

- Hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3$
 $H_a : \text{at least one population mean is different from the others}$
- $p - \text{value} = 0.5437$
- Conclusion: At the 5% significance level, there is enough evidence to suggest that the mean score on the final exam is the same for the different course delivery methods.

11. A local ski resort wants to know if the mean number of daily visitors is the three types of snow conditions. A sample of days is taken, and the number of visitors each day and the snow conditions are recorded. The results are shown in the table below. Assume the number of daily visitors are normally distribution and has equal variance.

Powder	Machine Made	Hard Packed
1,210	2,107	2,846
1,080	1,149	1,638
1,537	862	2,019
941	1,870	1,178
	1,528	2,233
	1,382	

At the 5\% significance level, determine if there is a difference in the mean number of daily visitors for the different snow conditions.

Click to see Answer

- Hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3$
 $H_a : \text{at least one population mean is different from the others}$
- $p - \text{value} = 0.0807$
- Conclusion: At the 5\% significance level, there is enough evidence to suggest that the mean number of daily visitors is the same for the different snow conditions.

“11.4 One-Way ANOVA and Hypothesis Tests for Three or More Population Means” and “11.5 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

PART XII

SIMPLE LINEAR REGRESSION

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it? In another example, the amount you pay a repair person for labour is often determined by an initial amount plus an hourly fee.

In this chapter, we will be studying the simplest form of regression analysis, **simple linear regression**, with one independent variable x . This involves data that fits a line in two dimensions. We will also study correlation, which measures the strength of the linear relationship.

CHAPTER OUTLINE

12.1 Linear Equations

12.2 Scatter Diagrams

12.3 Correlation

12.4 The Regression Equation

12.5 Coefficient of Determination

12.6 Standard Error of the Estimate

“12.1 Introduction to Linear Regression and Correlation” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

12.1 LINEAR EQUATIONS

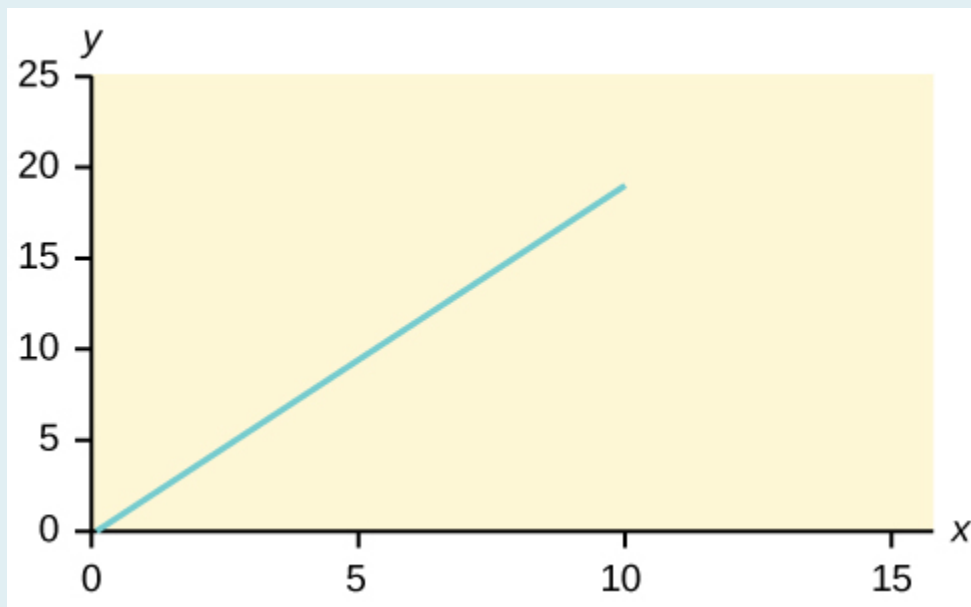
LEARNING OBJECTIVES

- Identify a linear equation, graphically or algebraically.

In this chapter, we will be studying simple linear regression, which models the linear relationship between two variables x and y . A linear equation has the form $y = b_0 + b_1x$ where b_0 is the y -intercept of the line and b_1 is the slope of the line. For example, $y = 3 + 2x$ and $y = 1 - 4x$ are examples of linear equations. The graph of linear equation is a straight line.

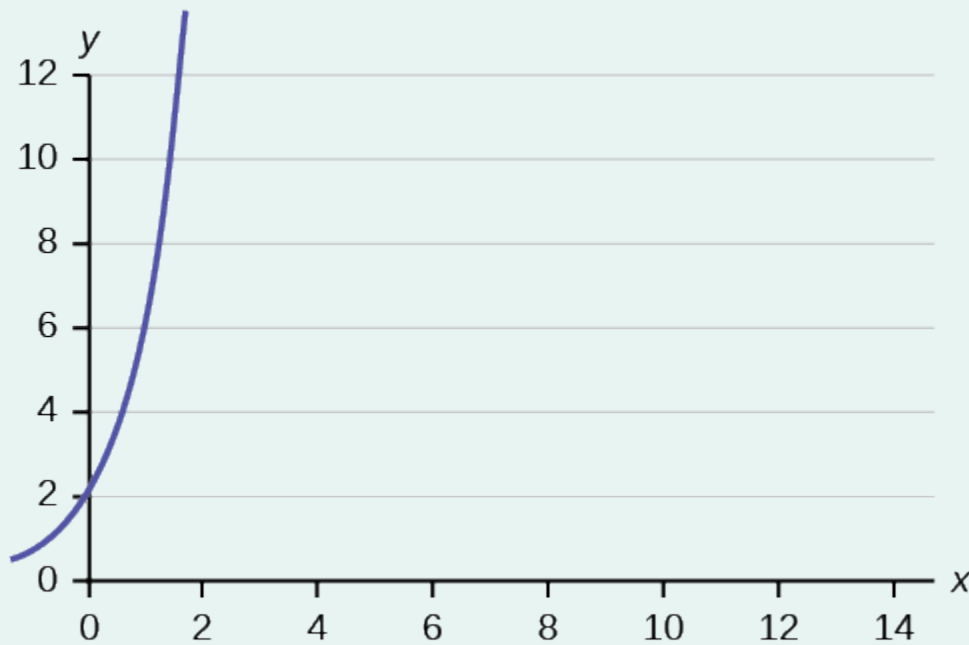
EXAMPLE

The equation $y = -1 + 2x$ is a linear equation. The slope is 2 and the y -intercept is -1 . The graph of $y = -1 + 2x$ is shown below.



TRY IT

Is the graph shown below the graph of a linear equation? Why or why not?



Click to see Solution

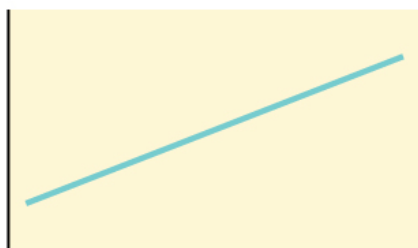
This is not a linear equation because the graph is not a straight line.

The **slope** b_1 is a number that describes the **steepness** of a line. The slope tells us how the value of the y variable will change for every one-unit increase in the value of the x variable.

The **y -intercept** b_0 is the value of the y -coordinate where the graph of the line crosses the y -axis. Algebraically, the y -intercept is the value of y when $x = 0$.

Consider the figure below, which illustrates three different linear equations:

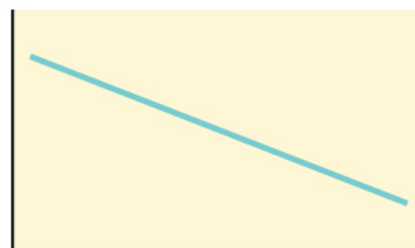
- In (a), the line rises from left to right across the graph. This means that the slope b_1 is a positive number ($b_1 > 0$).
- In (b), the line is horizontal (parallel to the x -axis). This means that the slope b_1 is zero ($b_1 = 0$).
- In (c), the line falls from left to right across the graph. This means that the slope b_1 is a negative number ($b_1 < 0$).



(a)



(b)



(c)

EXAMPLE

Consider the linear equation $y = -25 + 15x$.

- The slope is 15. This tells us that when the value of x increases by 1, the value of y increases by 15. Because the slope is positive, the graph of $y = -25 + 15x$ rises from left to right.
- The y -intercept is -25 . This tells us that when $x = 0$, $y = -25$. On the graph of $y = -25 + 15x$, the line crosses the y -axis at -25 .

TRY IT

Consider the linear equation $y = 17 - 10x$. Identify the slope and y -intercept. Describe the slope and y -intercept in sentences.

Click to see Solution

- The slope is -10 . This tells us that when the value of x increases by 1, the value of y decreases by 10. Because the slope is negative, the graph of $y = 17 - 10x$ falls from left to

right.

- The y -intercept is 17. This tells us that when $x = 0$, $y = 17$. On the graph of $y = 17 - 10x$, the line crosses the y -axis at 17.

Exercises

1. Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?

Click to see Answer

Not linear.

2. Which of the following equations are linear?

- a. $y = 6x + 8$
- b. $y + 7 = 3x$
- c. $y - x = 8x^2$
- d. $4y = 8$

Click to see Answer

(a), (b), and (d) are linear.

3. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15 - 1.5x$ where x is the number of hours passed in an eight-hour day of trading.
 - a. What is the slope? Interpret the slope's meaning in the context of the question.
 - b. What is the y -intercept? Interpret the y -intercept's meaning in the context of the question.
 - c. If you owned this stock, would you want a positive or negative slope? Why?

Click to see Answer

- a. -1.5 . For each additional hour that passes during the trading day, the price of the stock decreases by \$1.50.
- b. 15. At the start of the trading day, the price of the stock is \$15.

- c. Positive slope because that means the price of the stock is increasing.

“12.2 Linear Equations” and “12.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

12.2 SCATTER DIAGRAMS

LEARNING OBJECTIVES

- Define independent and dependent variables.
- Create and analyze scatter diagrams.

Independent and Dependent Variables

An **independent variable** (or the x -variable) is called the **explanatory** or **predictor** variable. The independent variable is used for prediction and provides the basis for estimation. The independent variable may be thought of as the input value and is used to determine the output value (the value of the dependent variable).

A **dependent variable** (or the y -variable) is called the **response** or **outcome** variable. The dependent variable is the variable being predicted or estimated based on the value of the independent variable. The dependent variable may be thought of as the output value and is determined by the input value (the value of the independent variable).

EXAMPLE

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time

fee of \$25 plus \$25 per hour of tutoring. The two variables are the number of hours per session and the amount of money earned per session.

- The number of hours per session is the independent variable because it can be used to predict the value of the other variable (the amount of money earned per session).
- The amount of money earned per session is the dependent variable because its value can be determined from the value of the other variable (the number of hours per session).

Scatter Diagrams

Before we begin discussing correlation and linear regression, we need to consider ways to display the relationship between the independent variable x and the dependent variable y . The most common and easiest way to illustrate the relationship between the two variables is with a scatter diagram.

A **scatter diagram** (or scatter plot) is a graphical presentation of the relationship between two numerical variables. Each point on the scatter diagram represents the values of two variables. The x -coordinate is the value of the independent variable, and the y -coordinate is the value of the corresponding dependent variable.

To construct a scatter diagram:

1. Identify the independent and dependent variables.
2. Assign the independent variable to the horizontal or x -axis. Assign the dependent variable to the vertical or y -axis.
3. Plot the points on an (x, y) -grid.
4. Label the axes, including both the variable names and units.
5. Include a chart title. A common chart title is ***independent variable vs dependent variable***, using the actual names of the variables.

EXAMPLE

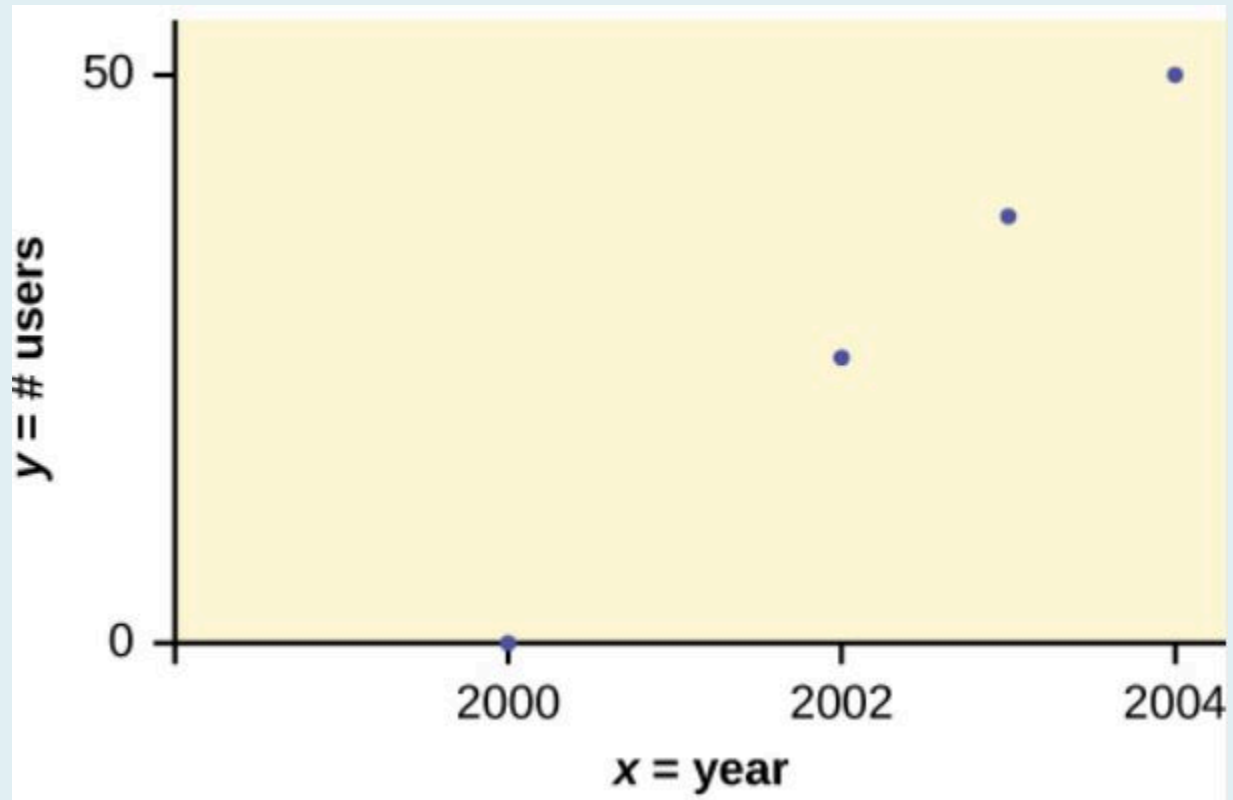
In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the internet. Data for the number of users from years 2000 through 2004 is given in the table below.

Year	Number of Users (in millions)
2000	0.5
2002	20.0
2003	33.0
2004	47.0

Which variable is the independent variable? Which variable is the dependent variable? Construct a scatter diagram for this data.

Solution

- The year is the independent variable because it can be used to predict the value of the other variable (the number of users).
- The number of users is the dependent variable because its value can be determined from the value of the other variable (year).



TRY IT

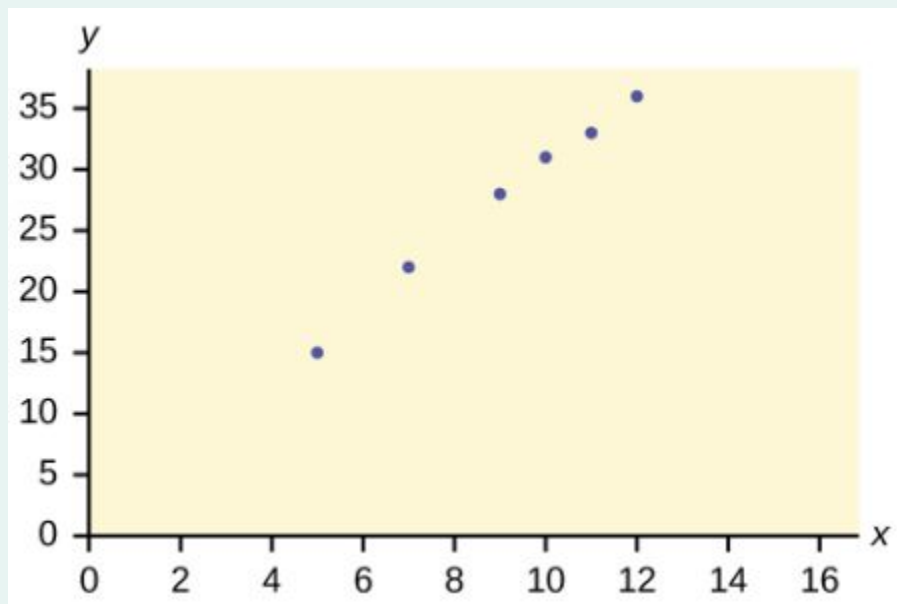
Amelia plays basketball for her high school. She wants to improve her play so she can compete at the college level. The table below records the number of hours she spends practicing her jump shot before a game and the number of points she scored in the following game.

Hours Spent Practicing Jump Shot	Points Scored in Game
5	15
7	22
9	28
10	31
11	33
12	36

Which variable is the independent variable? Which variable is the dependent variable? Construct a scatter diagram for this data.

Click to see Solution

- The hours spent practicing jump shots is the independent variable because it can be used to predict the value of the other variable (points scored in the game).
- The points scored in a game are the dependent variable because its value can be determined from the value of the other variable (hours spent practicing jump shots).



CONSTRUCTING A SCATTER DIAGRAM IN EXCEL

To create a scatter diagram in Excel:

1. Identify the independent and dependent variables.
2. If necessary, rearrange the columns so that the column containing the independent variable data is on the left and the dependent variable is on the right. (Excel always places the variable on the left on the horizontal axis.)
3. Go to the **Insert** tab. In the **Charts** group, click on **Scatter**. Select the scatter diagram with only markers (points).
4. Using the chart tools, add axis titles, including both the variable names and units on the axes.
5. Using the chart tools, add a chart title. A common chart title is ***independent variable vs dependent variable***, using the actual names of the variables.

Visit the Microsoft page for more information about creating a scatter diagram in Excel.

A scatter diagram shows the **direction** of the relationship between the independent and dependent variables. That is, a scatter diagram shows if the points are, in general, rising or falling as we read from left to right across the graph.

We can determine the **strength** of the relationship by looking at the scatter diagram to see how close the points are to a line, a power function, an exponential function, or to some other type of function. The stronger the relationship, the better the corresponding regression model (linear, exponential, etc.) will be at predicting values of the dependent variable.

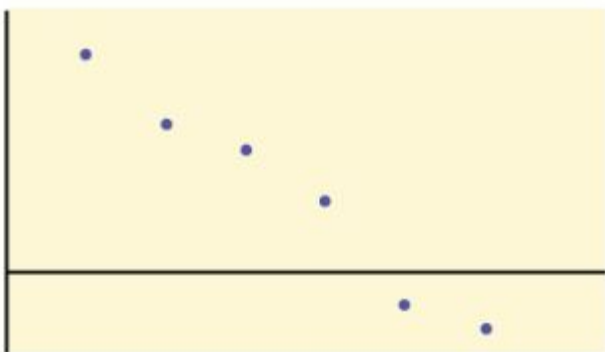
When we look at a scatter diagram, we want to notice the **overall pattern** and any **deviations** from the pattern. The scatter diagrams shown below illustrate these concepts.



(a) Positive linear pattern (strong)



(b) Linear pattern w/ one deviation



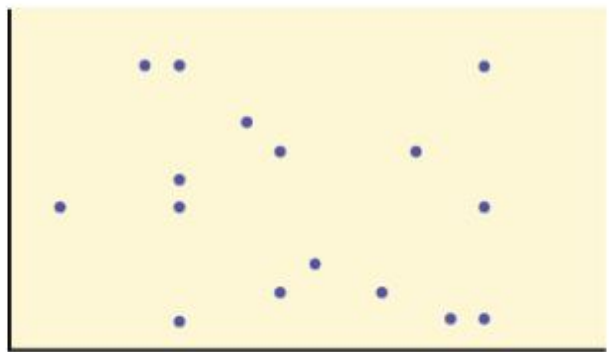
(a) Negative linear pattern (strong)



(b) Negative linear pattern (weak)



(a) Exponential growth pattern



(b) No pattern

In this chapter, we are only concerned with the strength and direction of the **linear** relationship between the independent and dependent variables. In the next section, we will learn about a numerical measure, the correlation coefficient, that measures the strength and direction of the linear relationship.

Because linear patterns are quite common, we are interested in scatter diagrams that show a linear pattern. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If a scatter diagram shows a linear relationship, we would like to create a model based on this apparent linear relationship. This model is constructed through a process called **simple linear regression**. However, we only calculate a regression line if one of the variables, x , helps to explain or predict the other variable, y . If x is the independent variable and y is the dependent variable, then we can use a regression line to predict a value for y for a given value of x .



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=284#oembed-1>

Video: “Basic Excel Business Analytics #44: Intro To Linear Regression & Scatter Chart” by excelisfun [15:46] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

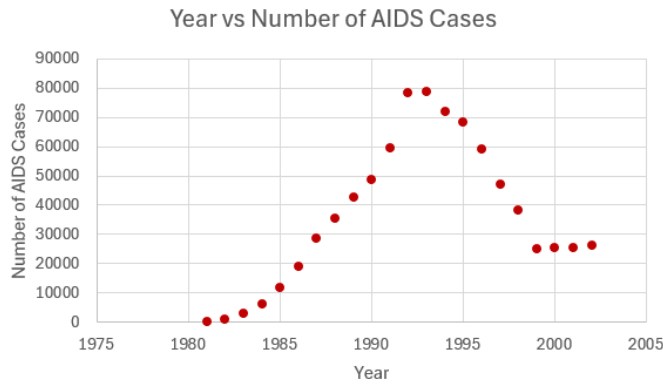
1. The table below contains real data for the first two decades of AIDS cases.

Year	Number of AIDS Cases
1981	319
1982	1,170
1983	3,076
1984	6,240
1985	11,776
1986	19,032
1987	28,564
1988	35,447
1989	42,674
1990	48,634
1991	59,660
1992	78,530
1993	78,834
1994	71,874
1995	68,505
1996	59,347
1997	47,149
1998	38,393
1999	25,174
2000	25,522
2001	25,643
2002	26,464

- Which variable is the independent variable, and which variable is the dependent variable?
- Construct a scatter diagram for this data.

Click to see Answer

- Independent: Year; Dependent: Number of AIDS Cases



b.

2. A specialty cleaning company charges an equipment fee and an hourly labour fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

- What are the independent and dependent variables?
- What is the y -intercept? Interpret the y -intercept using complete sentences.
- What is the slope? Interpret the slope using complete sentences.

Click to see Answer

- Independent: Number of Hours of Labour; Dependent: Total Amount of Fee
50. When the number of hours of labour is 0, the total amount is \$50.
100. For each extra hour of labour, the total amount increases by \$100.

3. Due to erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12,000x$.
- What are the independent and dependent variables?
 - How many pounds of soil does the shoreline lose in a year?
 - What is the y -intercept? Interpret its meaning.

Click to see Answer

- Independent: Year; Dependent: Amount of Soil Lost
- 12,000
0. There is no soil lost in year 0.

4. For each of the following situations, state the independent variable and the dependent variable.
- A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared to

the age of drivers.

- b. A study is done to determine if the weekly grocery bill changes based on the number of family members.
- c. Insurance companies base life insurance premiums partially on the age of the applicant.
- d. Utility bills vary according to power consumption.
- e. A study is done to determine if a higher education reduces the crime rate in a population.

Click to see Answer

- a. Independent: Age of driver; Dependent: Number of fatalities
- b. Independent: Number of family members; Dependent: Weekly grocery bill
- c. Independent: Age of applicant; Dependent: Life insurance premium
- d. Independent: Power consumption; Dependent: Amount of utility bill
- e. Independent: Years of education; Dependent: Crime rate

5. The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. The table below shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Year	Cuba's PPP
1999	1,700
2000	1,700
2002	2,300
2003	2,900
2004	3,000
2005	3,500
2006	4,000
2007	11,000
2008	9,500
2009	9,700
2010	9,900

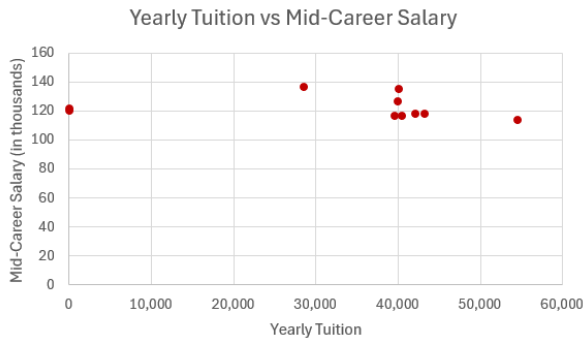
Click to see Answer



6. Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data.

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Click to see Answer



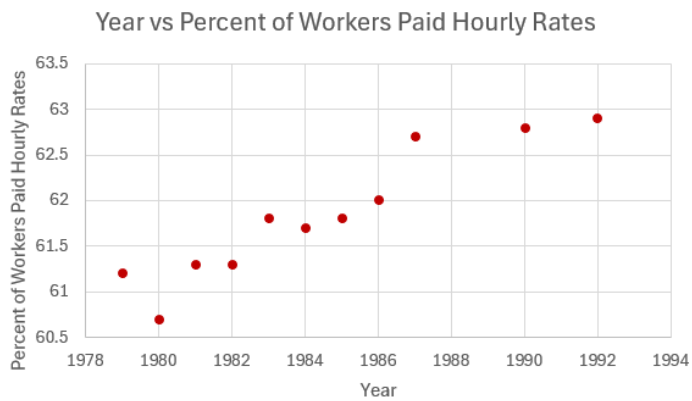
7. The table below gives the percent of workers who are paid hourly rates for the years 1979 to 1992.

Year	Percent of Workers Paid Hourly Rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

- Decide which variable should be the independent variable and which should be the dependent variable.
- Draw a scatter plot of the ordered pairs.

Click to see Answer

- Independent: year; Dependent: percent of workers paid hourly rate



b.

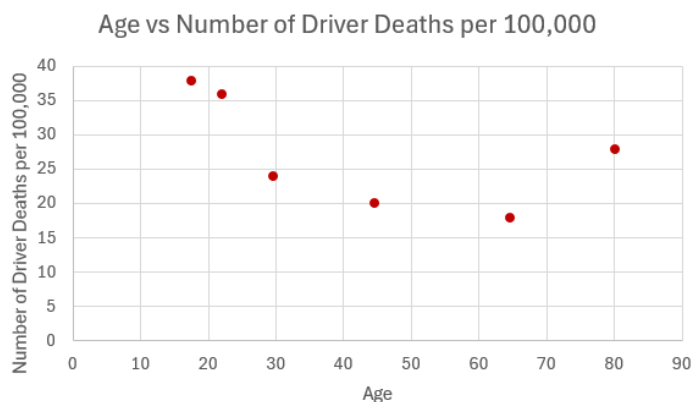
8. Recently, the annual number of driver deaths per 100, 000 for the selected age groups was as follows:

Age	Number of Driver Deaths per 100, 000
17.5	38
22	36
29.5	24
44.5	20
64.5	18
80	28

- Decide which variable should be the independent variable and which should be the dependent variable.
- Draw a scatter plot of the data.

Click to see Answer

- Independent: age; Dependent: number of driver deaths per 100, 000



b.

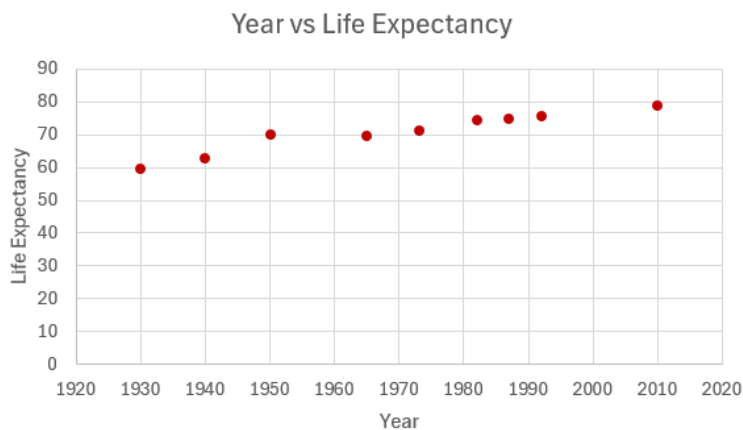
9. The table below shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

- Decide which variable should be the independent variable and which should be the dependent variable.
- Draw a scatter plot of the ordered pairs.

Click to see Answer

- Independent: year; Dependent: life expectancy



b.

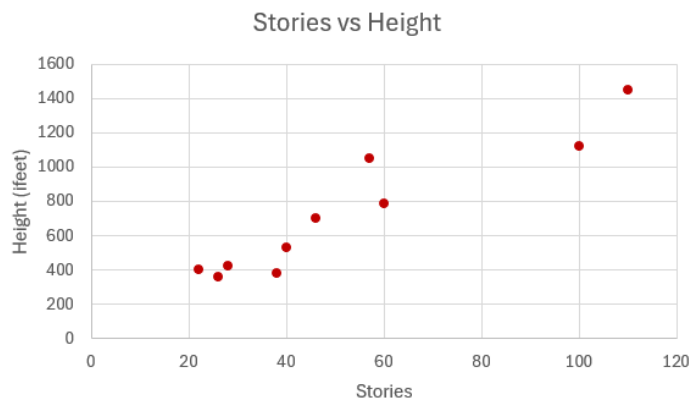
10. The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level).

Height (in feet)	Number of Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

- Decide which variable should be the independent variable and which should be the dependent variable.
- Draw the scatter diagram for this data.

Click to see Answer

- Independent: number of stories; Dependent: height



b.

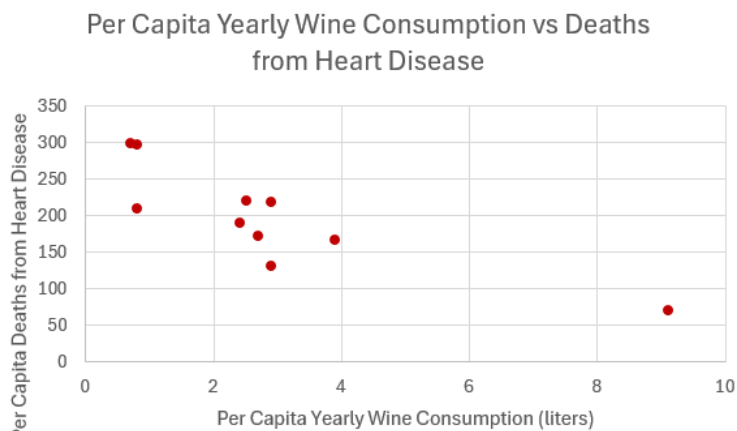
- The following table shows data on average per capita wine consumption and heart disease rate in a random sample of 10 countries.

Per Capita Yearly Wine Consumption in Liters	Per Capita Death from Heart Disease
2.5	221
3.9	167
2.9	131
2.4	191
2.9	220
0.8	297
9.1	71
2.7	172
0.8	211
0.7	300

- Decide which variable should be the independent variable and which should be the dependent variable.
- Draw a scatter plot of the ordered pairs.

Click to see Answer

- Independent: yearly wine consumption; Dependent: death from heart disease



b.

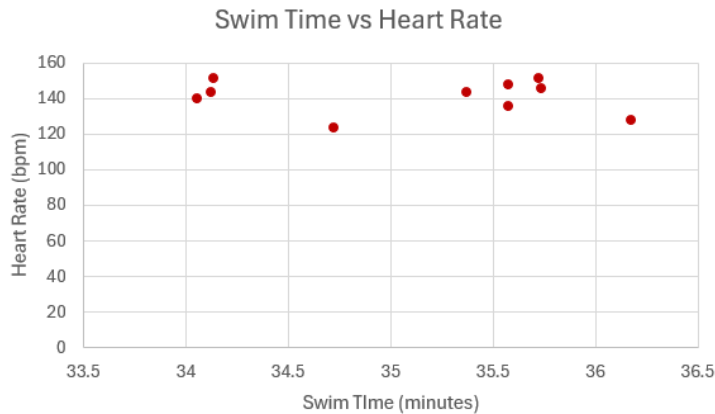
- The following table consists of one student athlete's time (in minutes) to swim 2000 meters and the student's heart rate (beats per minute) after swimming on a random sample of 10 days.

Swim Time	Heart Rate
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

- Decide which variable should be the independent variable and which should be the dependent variable.
- Draw a scatter plot of the ordered pairs.

Click to see Answer

- Independent: swim time; Dependent: heart rate



b.

“12.3 Scatter Diagrams” and “12.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

12.3 CORRELATION

LEARNING OBJECTIVES

- Calculate and interpret the correlation coefficient.

The purpose of simple linear regression is to build a linear model that can be used to make predictions of the y variable for a given value of the x variable. Of course, we want the model to give us good predictions—there is no point in using a model that gives bad or inaccurate predictions. But how can we tell if the linear model will provide accurate predictions? As we have seen, we can examine the scatter diagram for a set of data to get a sense of the strength and direction of the linear relationship between the independent variable x and the dependent variable y . But we would like a **numerical measure** of the strength and direction of the linear relationship we observe on the scatter diagram. This numerical measure is called the **correlation coefficient**.

The correlation coefficient was developed by Karl Pearson in the early 1900s and is sometimes referred to as Pearson's correlation coefficient. Denoted by r , the **correlation coefficient** is a numerical measure of the strength and direction of the linear relationship between the independent variable x and the dependent variable y . Although there is an algebraic formula to find the value of r , we will perform the calculation using the built-in function in Excel.

Interpreting the Correlation Coefficient

What does the value of r tell us?

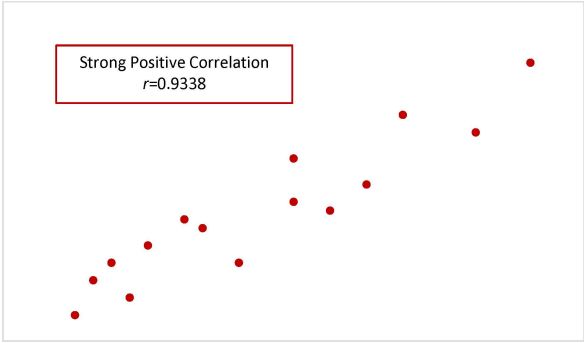
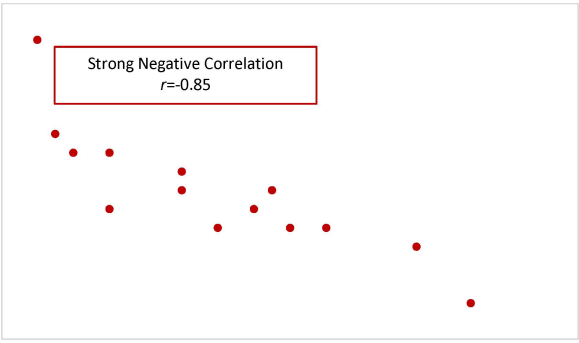
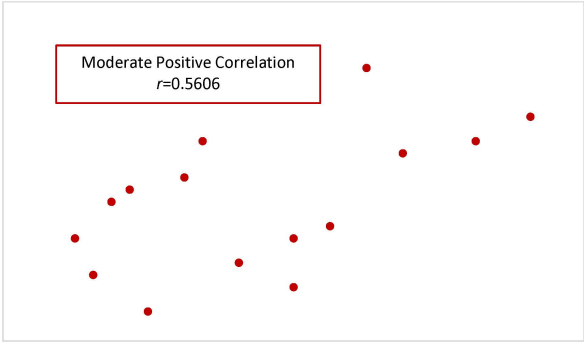
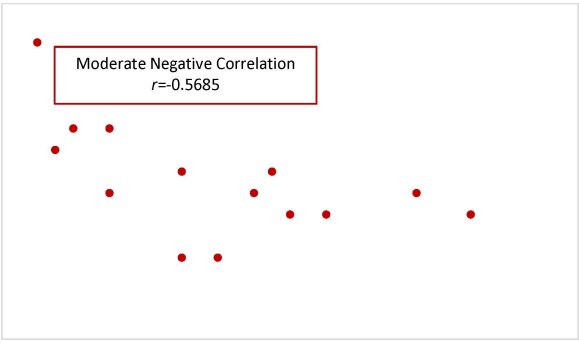
- The value of the correlation coefficient r is always a number between -1 and 1 .
- Values of r close to 1 or -1 indicate a **strong** linear relationship between x and y . If $r = 1$, then there is a perfect, positive correlation between x and y , in which case the points on the

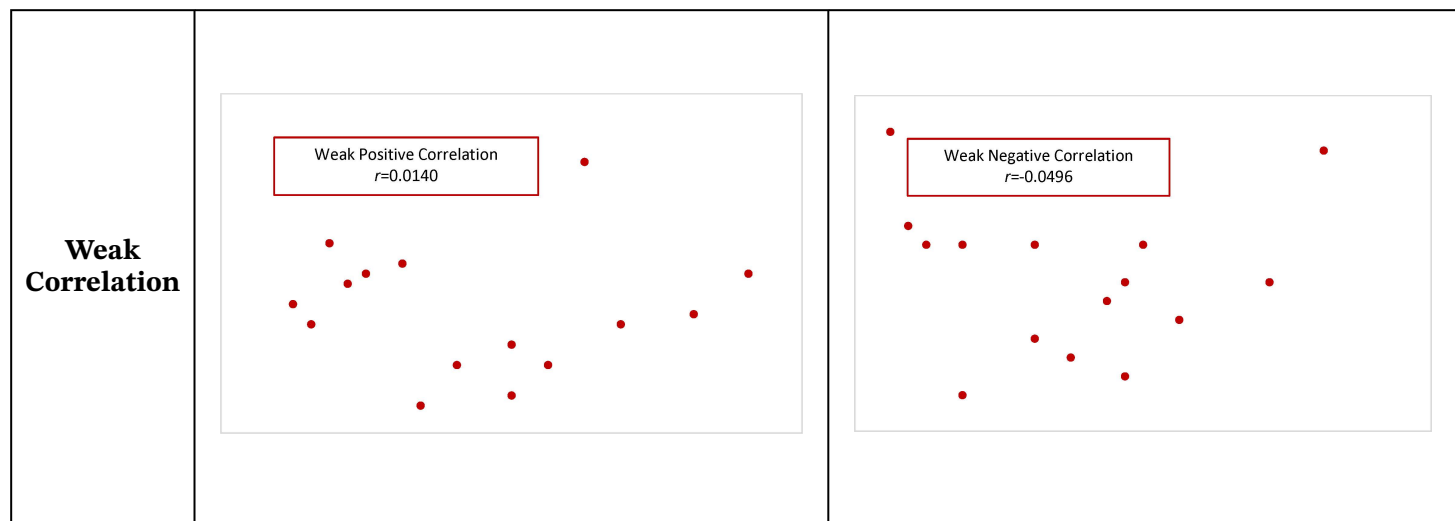
scatter diagram would all lie on a straight line that rises from left to right. If $r = -1$, then there is a perfect, negative correlation between x and y , in which case the points on the scatter diagram would all line on a straight line that falls from left to right.

- Values of r close to 0.5 or -0.5 indicate a **moderate** linear relationship between x and y .
- Values of r close to 0 indicate a **weak** linear relationship between x and y . If $r = 0$, then there is no correlation between x and y .

What does the sign of r tell us?

- A positive value of r means that the points on the scatter diagram have the general tendency to rise from left to right. In other words, when x increases, y tends to increase, and when x decreases, y tends to decrease.
- A negative value of r means that the points on the scatter diagram have the general tendency to fall from left to right. In other words, when x increases, y tends to decrease, and when x decreases, y tends to increase.

	<p>Positive Correlation</p>	<p>Negative Correlation</p>
<p>Strong Correlation</p>	 <p>Strong Positive Correlation $r=0.9338$</p> <p>A scatter plot with 15 red data points showing a strong positive linear relationship. The points are tightly clustered along a diagonal line sloping upwards from left to right. A red-bordered text box in the upper-left corner of the plot area contains the text 'Strong Positive Correlation' and '$r=0.9338$'.</p>	 <p>Strong Negative Correlation $r=-0.85$</p> <p>A scatter plot with 15 red data points showing a strong negative linear relationship. The points are tightly clustered along a diagonal line sloping downwards from left to right. A red-bordered text box in the upper-left corner of the plot area contains the text 'Strong Negative Correlation' and '$r=-0.85$'.</p>
<p>Moderate Correlation</p>	 <p>Moderate Positive Correlation $r=0.5606$</p> <p>A scatter plot with 15 red data points showing a moderate positive linear relationship. The points are more spread out than in the strong correlation plot but still follow a clear upward trend. A red-bordered text box in the upper-left corner of the plot area contains the text 'Moderate Positive Correlation' and '$r=0.5606$'.</p>	 <p>Moderate Negative Correlation $r=-0.5685$</p> <p>A scatter plot with 15 red data points showing a moderate negative linear relationship. The points are more spread out than in the strong correlation plot but still follow a clear downward trend. A red-bordered text box in the upper-left corner of the plot area contains the text 'Moderate Negative Correlation' and '$r=-0.5685$'.</p>



CALCULATING THE CORRELATION COEFFICIENT IN EXCEL

To calculate the correlation coefficient, use the **correl(array,array)** function. Enter the cell array containing the independent variable data into one of the arrays and enter the cell array containing the dependent variable data into the other array.

The output from the **correl** function is the value of the correlation coefficient.

Visit the Microsoft page for more information about the **correl** function.

NOTE

The arrays containing the independent and dependent variable data may be entered into the **correl** function in either order. The output from the **correl** function does not depend on the order in which the arrays are entered.

EXAMPLE

A statistics professor wants to study the relationship between a student's score on the third exam in the course and their final exam score. The professor took a random sample of 11 students and recorded their third exam score (out of 80) and their final exam score (out of 200). The results are recorded in the table below.

Student	Third Exam Score	Final Exam Score
1	65	175
2	67	133
3	71	185
4	71	163
5	66	126
6	75	198
7	67	153
8	70	163
9	71	159
10	69	151
11	69	159

1. Find the correlation coefficient for this data.
2. Interpret the correlation coefficient found in part 1.

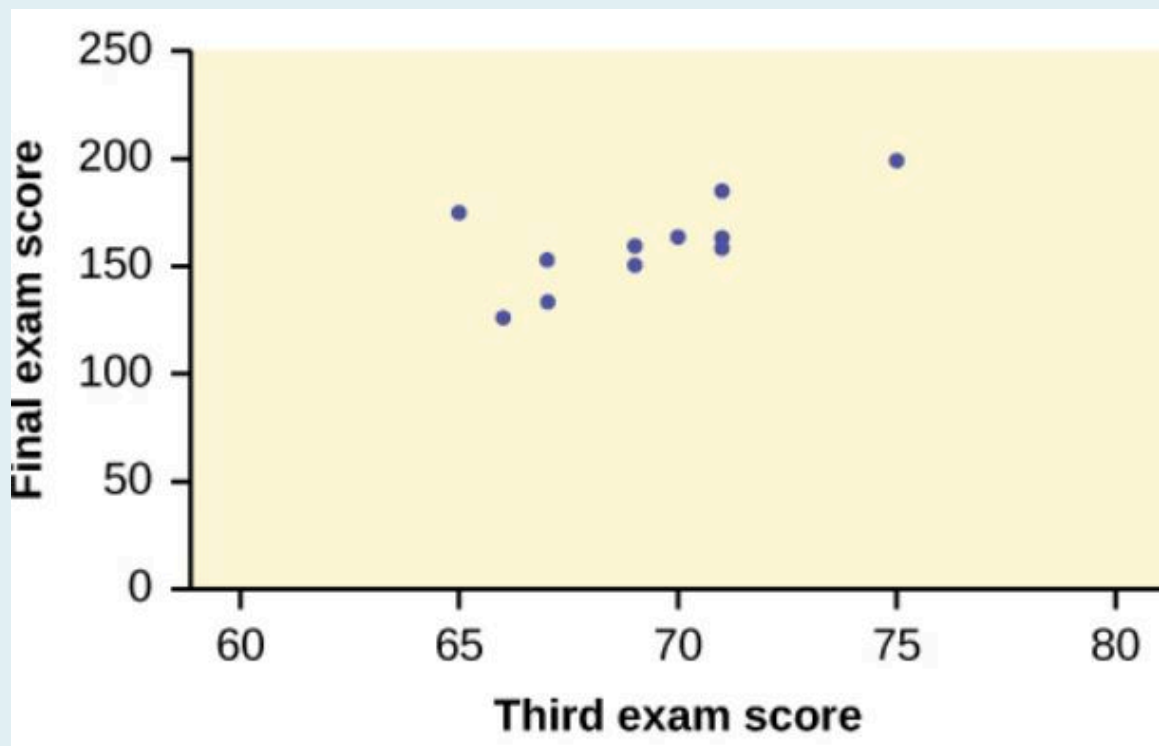
Solution

1. Enter the data into an Excel spreadsheet. For this example, suppose we entered the data (without the column headings) so that the student column is in column A from A1 to A11, the third exam score is in column B from B1 to B11, and the final exam score is in column C from C1 to C11.

Function	correl
Field 1	B1:B11
Field 2	C1:C11
Answer	0.6631

The value of the correlation coefficient is $r = 0.6631$.

By examining the scatter diagram for this data, shown below, we can see that the points are rising from left to right (corresponding to the fact that r is positive) and the general pattern of points corresponds to a moderate linear relationship (corresponding to the fact that r is close to 0.5).



- There is a moderate, positive linear relationship between the third test score and the final exam score.

NOTES

1. In this case, the value of r is close to 0.5 , so we would consider this a moderate linear relationship.
2. When writing down the interpretation of the correlation coefficient, remember to be specific to the question using the actual names of the independent and dependent variables. Also make sure to include in the sentence the strength of the linear relationship (strong, moderate, or weak) and the direction of the linear relationship (positive or negative).

TRY IT

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the table below shows different depths with the maximum dive times in minutes.

Depth (in feet)	Maximum Dive Time (in minutes)
50	80
60	55
70	45
80	35
90	25
100	22

1. Find the correlation coefficient for this data.
2. Interpret the correlation coefficient found in part 1.

Click to see Solution

1. $r = -0.9629$
2. There is a strong, negative linear relationship between depth and maximum dive time.

Correlation versus Causation

The correlation coefficient only measures the **correlation** between two variables, not **causation**. A strong correlation between two variables does not mean that changes in one variable actually cause changes in the other variable. The correlation coefficient can only tell us that changes in the independent variable and dependent variable are related. In general, remember that “correlation does not equal causation.”

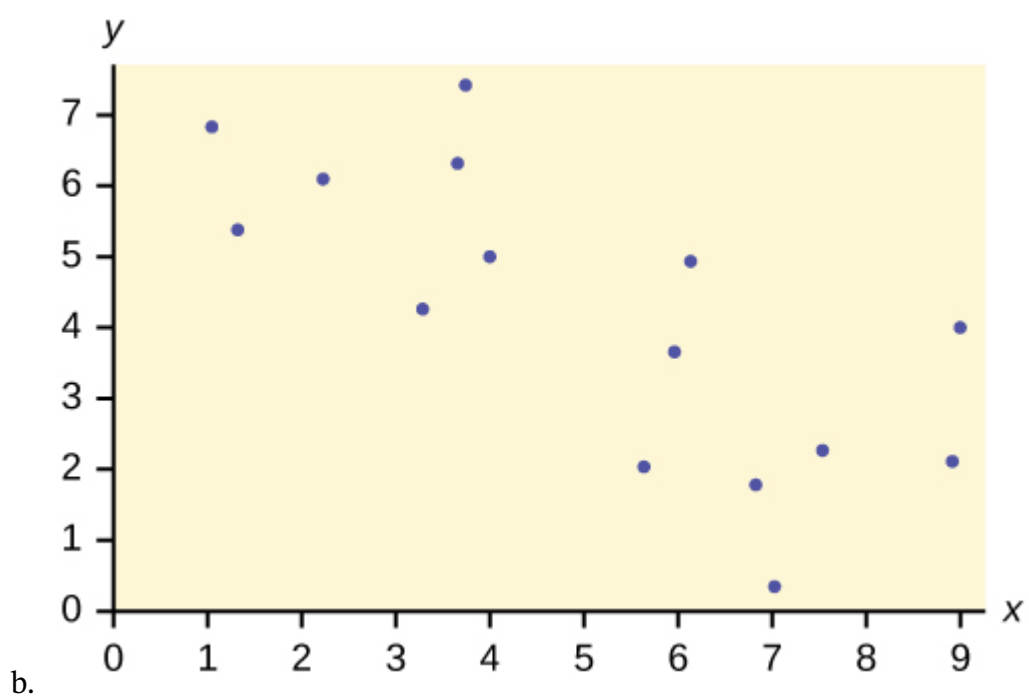
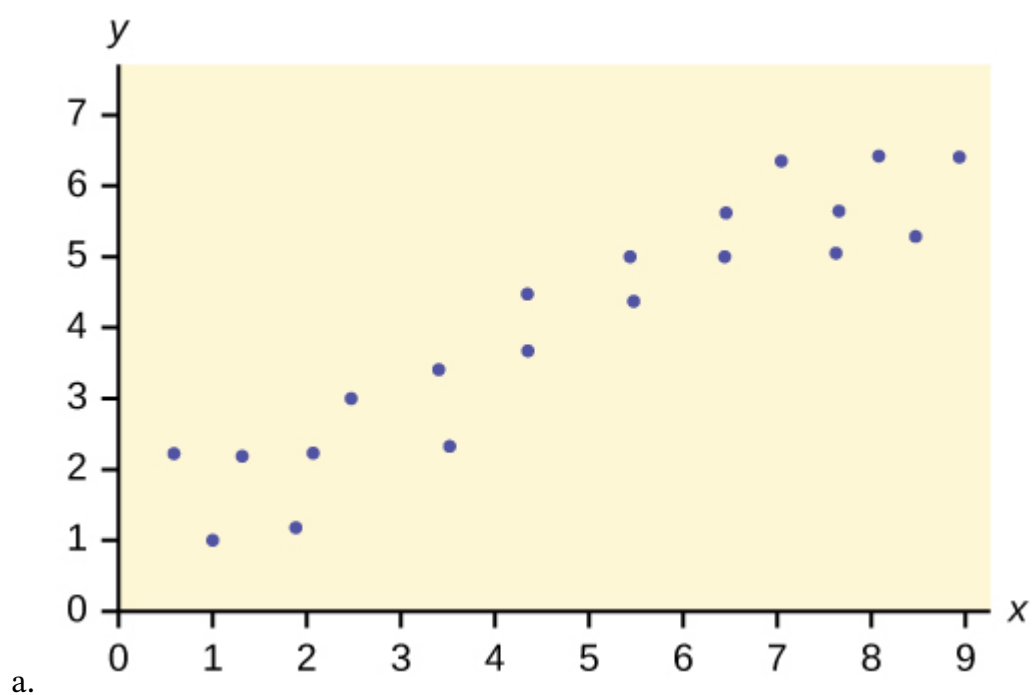


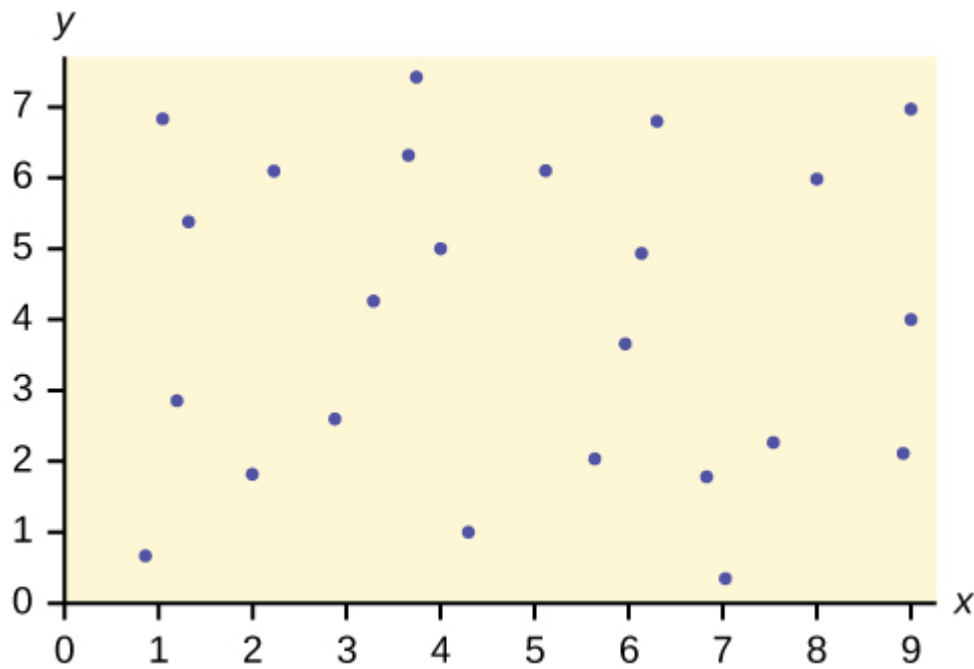
One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=293#oembed-1>

Video: “Using Excel to calculate a correlation coefficient || interpret relationship between variables” by Matt Macarty [5:22] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. For each scatter diagram shown below, determine if the value of the correlation coefficient indicates a strong, moderate, or weak linear relationship and a positive or negative linear relationship.





c.

Click to see Answer

- a. strong, positive
- b. moderate, negative
- c. weak, positive

2. What does an r value of 0 mean?**Click to see Answer**

There is no linear relationship between the independent and dependent variables.

3. Interpret each of the following values of r .

- a. $r = 0.67$
- b. $r = -0.12$
- c. $r = -0.93$

Click to see Answer

- a. There is a moderate, positive linear relationship between the independent and dependent variables.
- b. There is a weak, negative linear relationship between the independent and dependent variables.
- c. There is a strong, negative linear relationship between the independent and dependent variables.

4. In a random sample of ten professional athletes, the number of endorsements the player has and the amount of money (in millions of dollars) the player earns are recorded in the table below.

Player	Number of Endorsements	Money Earned (in millions)
1	0	2
2	3	8
3	2	7
4	1	3
5	5	13
6	5	12
7	4	9
8	3	9
9	0	3
10	4	10

- Calculate the correlation coefficient.
- Interpret the correlation coefficient.

Click to see Answer

- 0.9786
- There is a strong, positive linear relationship between the number of endorsements and money earned.

5. The table below gives the percentage of workers who are paid hourly rates for the years 1979 to 1992.

Year	Percent of Workers Paid Hourly Rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

- a. Find the correlation coefficient
- b. Interpret the correlation coefficient.

Click to see Answer

- a. 0.9448
- b. There is a strong, positive linear relationship between the year and the percentage of workers paid an hourly rate.

6. The table below contains real data for the first two decades of AIDS cases.

Year	Number of AIDS Cases
1981	319
1982	1,170
1983	3,076
1984	6,240
1985	11,776
1986	19,032
1987	28,564
1988	35,447
1989	42,674
1990	48,634
1991	59,660
1992	78,530
1993	78,834
1994	71,874
1995	68,505
1996	59,347
1997	47,149
1998	38,393
1999	25,174
2000	25,522
2001	25,643
2002	26,464

- Calculate the correlation coefficient.
- Interpret the correlation coefficient.

Click to see Answer

- 0.4526
- There is a moderate, positive linear relationship between the year and the number of AIDS cases.

7. Recently, the annual number of driver deaths per 100, 000 for the selected age groups was as follows:

Age	Number of Driver Deaths per 100, 000
17.5	38
22	36
29.5	24
44.5	20
64.5	18
80	28

- Find the correlation coefficient.
- Interpret the correlation coefficient.

Click to see Answer

- 0.5787
 - There is a moderate, negative linear relationship between age and the number of driver deaths per 100, 000.
8. The table below shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

- Find the correlation coefficient

- b. Interpret the correlation coefficient.

Click to see Answer

- a. 0.9614
b. There is a strong, positive linear relationship between year and life expectancy.

9. The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level).

Height (in feet)	Number of Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

- a. Find the correlation coefficient
b. Interpret the correlation coefficient.

Click to see Answer

- a. 0.9436
b. There is a strong, positive linear relationship between number of stories and height.

10. The following table shows data on average per capita wine consumption and heart disease rate in a random sample of 10 countries.

Per Capita Yearly Wine Consumption in Liters	Per Capita Death from Heart Disease
2.5	221
3.9	167
2.9	131
2.4	191
2.9	220
0.8	297
9.1	71
2.7	172
0.8	211
0.7	300

- Find the correlation coefficient
- Interpret the correlation coefficient.

Click to see Answer

- −0.8359
- There is a strong, negative linear relationship between per capita yearly wine consumption and per capita deaths from heart disease.

- The following table consists of one student athlete's time (in minutes) to swim 2000 meters and the student's heart rate (beats per minute) after swimming on a random sample of 10 days.

Swim Time	Heart Rate
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

- Find the correlation coefficient
- Interpret the correlation coefficient.

Click to see Answer

- −0.1236
- There is a weak, negative linear relationship between swim time and heart rate.

“12.4 Correlation” and “12.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

12.4 THE REGRESSION EQUATION

LEARNING OBJECTIVES

- Find the equation of the line-of-best fit.
- Use the line-of-best-fit to make predictions.

We often want to use the values of the independent variable to make predictions about the value of the dependent variable. For example, we might want to use the amount a business spends on advertising each quarter to make a prediction about the revenue the business will generate that quarter. When a linear relationship exists between an independent and dependent variable, we can build a linear model of that relationship, and then we can use that model to make predictions about the dependent variable.

Simple linear regression is a modelling technique in which the linear relationship between one independent variable x and one dependent variable y is approximated by a straight line, called the **line-of-best-fit** or **least squares line**. It is important to note that the line-of-best-fit only models the linear relationship between the independent and dependent variables.

The equation for the regression line is

$$\hat{y} = b_0 + b_1x$$

\hat{y} = predicted value of y

x = value of the independent variable

b_0 = y -intercept of the line

b_1 = slope of the line

The value of \hat{y} is the **estimated value of y** . It is the value of y obtained using the regression line. The value of \hat{y} is not generally equal to the value of y from the sample data. The values for the slope

b_1 and the y -intercept b_0 in the line-of-best-fit are calculated using the sample data and the **least squares method**. Although there are formulas to calculate the values of the slope and y -intercept in the regression line, we will calculate the slope and y -intercept using the built-in functions in Excel.

What does the slope of the linear regression equation tell us?

- The slope of the line-of-best-fit b_1 and the correlation coefficient r have the same sign. That is, b_1 and r are either both positive or both negative.
- The slope b_1 of the regression equation tells us how the dependent variable y changes for a one-unit increase in the independent variable x .
- When interpreting the slope, be specific to the context of the question, using the actual names of the variable and correct units.

What does the y -intercept of the linear regression equation tell us?

- The y -intercept b_0 of the line-of-best-fit is the predicted value of the dependent variable y when $x = 0$.
- When interpreting the y -intercept, be specific to the context of the question, using the actual names of the variable and correct units.

CALCULATING THE SLOPE AND y -INTERCEPT OF THE LINEAR REGRESSION EQUATION IN EXCEL

To calculate the slope of the linear regression equation, use the **slope(array for y's, array for x's)** function.

- For **array for y's**, enter the cell array containing the **dependent** variable y data.
- For **array for x's**, enter the cell array containing the **independent** variable x data.

Visit the Microsoft page for more information about the **slope** function.

To calculate the y -intercept of the linear regression equation, use the **intercept(array for y's, array for x's)** function.

- For **array for y's**, enter the cell array containing the **dependent** variable y data.
- For **array for x's**, enter the cell array containing the **independent** variable x data.

Visit the Microsoft page for more information about the **intercept** function.

NOTE

The order in which the data is entered into these functions is important. In both the slope and intercept functions, the data for the **dependent** variable is entered in the **first** array, and the data for the **independent** variable is entered in the **second** array. The output from the **slope** and **intercept** function will be different when the order of the inputs are switched.

EXAMPLE

A statistics professor wants to study the relationship between a student's score on the third exam in the course and their final exam score. The professor took a random sample of **11** students and recorded their third exam score (out of 80) and their final exam score (out of 200). The results are recorded in the table below. The professor wants to develop a linear regression model to predict a student's final exam score from the third exam score.

Student	Third Exam Score	Final Exam Score
1	65	175
2	67	133
3	71	185
4	71	163
5	66	126
6	75	198
7	67	153
8	70	163
9	71	159
10	69	151
11	69	159

1. Find the equation for the line-of-best-fit.
2. Interpret the slope of the line-of-best fit.

Solution

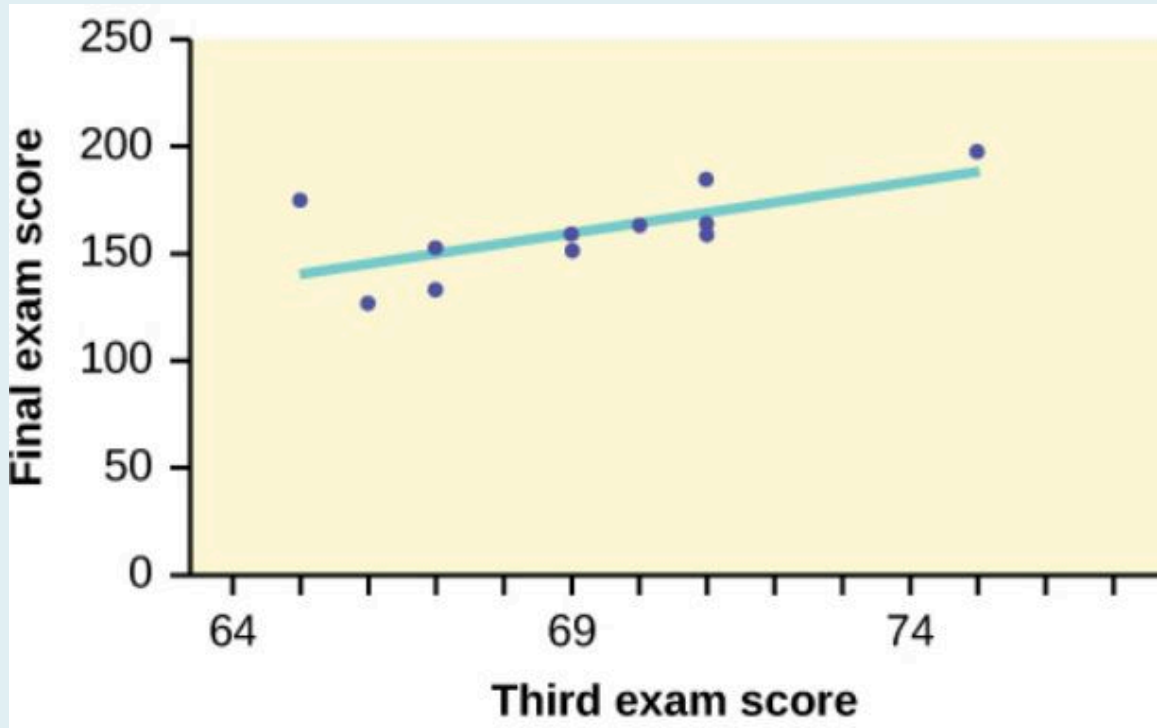
1. Because we want to predict the final exam score from the third exam score, the independent variable x is the third exam score, and the dependent variable y is the final exam score. Enter the data into an Excel spreadsheet. For this example, suppose we entered the data (without the column headings) so that the student column is in column A from A1 to A11, the third exam score is in column B from B1 to B11, and the final exam score is in column C from C1 to C11.

Function	slope
Field 1	C1:C11
Field 2	B1:B11
Answer	4.83

Function	intercept
Field 1	C1:C11
Field 2	B1:B11
Answer	-173.51

The equation for the line-of-best-fit is $\hat{y} = -173.51 + 4.83x$ where x is the third exam score and \hat{y} is the (predicted) final exam score.

The graph below shows the scatter diagram with the line-of-best fit.



2. The slope is $b_1 = 4.83$. Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points.

NOTE

1. When writing down the linear regression equation, remember to define what the variables represent in the context of the question. That is, state what x and \hat{y} represent in relation to the question.
2. When writing down the interpretation of the slope, remember to be specific to the question using the actual names of the independent and dependent variables and appropriate units.

Making Predictions with the Linear Regression Equation

Given a specific value of the independent variable x , the linear regression equation may be used to predict/estimate the value of the dependent variable y . To make predictions, the following conditions must be met:

- There must be a linear relationship between the variables. The stronger the linear relationship, the better the prediction will be.
- The linear regression equation is only valid to predict the values of the dependent variable. That is, we may only use the equation to solve for \hat{y} for a given value of x , and not the other way around.
- The linear regression equation should only be used to make predictions for y for values of x within the domain of the x values in the sample data used to construct the regression equation. The regression equation does not provide reliable predictions for values of x that fall outside the domain of the x values in the sample data.

EXAMPLE

A statistics professor wants to study the relationship between a student's score on the third exam in the course and their final exam score. The professor took a random sample of 11 students and recorded their third exam score (out of 80) and their final exam score (out of 200). The results are recorded in the table below. The professor developed the linear regression model

$\hat{y} = -173.51 + 4.83x$ to predict a student's final exam score (\hat{y}) from a student's third exam score (x).

Student	Third Exam Score	Final Exam Score
1	65	175
2	67	133
3	71	185
4	71	163
5	66	126
6	75	198
7	67	153
8	70	163
9	71	159
10	69	151
11	69	159

1. What is the professor's final exam prediction for a student who scored **66** on the third exam?
2. What is the professor's final exam prediction for a student who scored **73** on the third exam?
3. Should the professor use the linear regression model to predict the final exam score for a student who scored **80** on the third exam? Why?

Solution

1. Substitute $x = 66$ into the linear regression equation:

$$\begin{aligned}\hat{y} &= -173.51 + 4.83 \times 66 \\ &= 145.27\end{aligned}$$

A student who scored **66** on the third exam has a predicted score of **145.27** on the final exam.

2. Substitute $x = 73$ into the linear regression equation:

$$\begin{aligned}\hat{y} &= -173.51 + 4.83 \times 73 \\ &= 179.08\end{aligned}$$

A student who scored **73** on the third exam has a predicted score of **179.08** on the final exam.

3. The x values (third exam score) in the sample data are between **65** and **75**. An x value of **80**

is outside the domain of the observed x values in the data. So, we cannot **reliably** predict the final exam score for a student who scored 80 on the third exam. Of course, it is possible to enter $x = 80$ into the linear regression equation and calculate the corresponding value of \hat{y} , but this value is not a reliable prediction. If we calculate out the value of \hat{y} in the regression equation for $x = 80$, we get $\hat{y} = 212.89$, a value that makes no sense in the context of the question because the maximum score on the final exam is 200.

NOTES

1. The values obtained for the linear regression equation are predictions only. Here, **145.27** is the **predicted** final exam score for a student who scored **66** on the third exam. This does not mean that a student who actually scored **66** on the third exam will score **145.27** on the final exam.
2. Remember that the linear regression equation only gives reliable predictions for values of x that fall within the domain of x values in the sample data.

TRY IT

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the table below shows different depths with the maximum dive times in minutes.

Depth (in feet)	Maximum Dive Time (in minutes)
50	80
60	55
70	45
80	35
90	25
100	22

1. Find the linear regression equation to predict the maximum dive time from the depth.
2. Interpret the slope of the regression equation found in part 1.
3. Predict the maximum dive time for a depth of 75 feet.

Click to see Solution

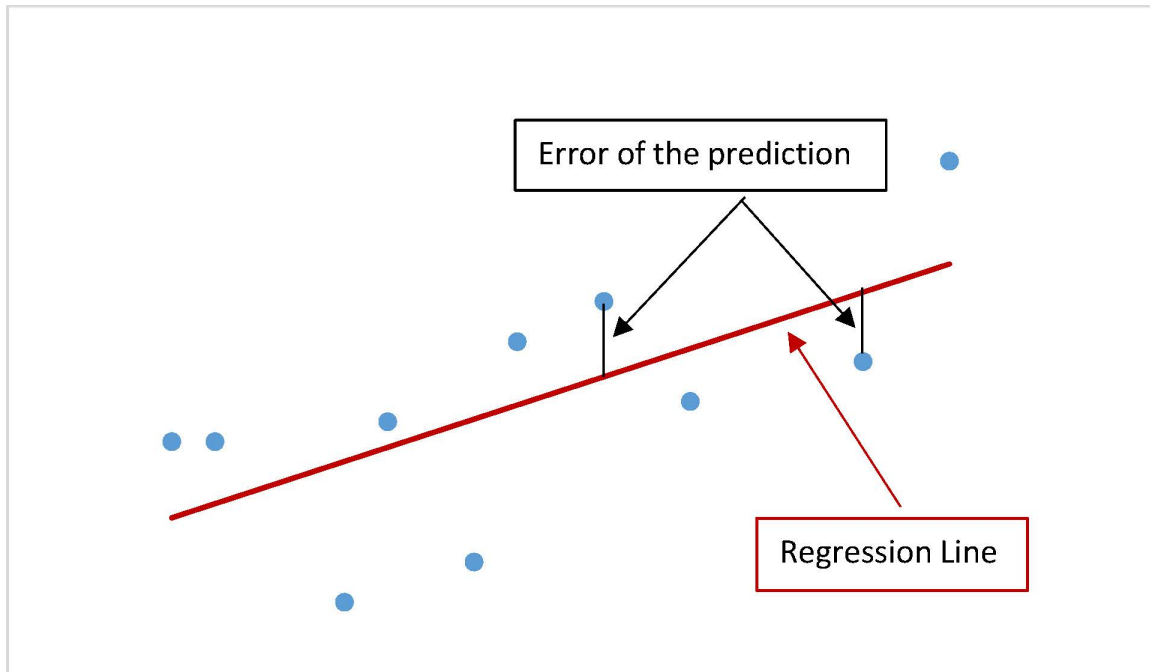
1. $\hat{y} = 127.24 - 1.11x$ where x is the depth in feet and \hat{y} is the (predicted) maximum dive time in minutes.
2. For each one-foot increase in depth, the maximum dive time decreases by 1.11 minutes.
3. $\hat{y} = 127.24 - 1.11 \times 75 = 43.99$ minutes

Errors and The Least Squares Method

The difference between the actual value of the dependent variable y (in the sample data) and the predicted value of the dependent variable \hat{y} obtained from the linear regression equation is called the **error** or **residual**.

$$\begin{aligned}\text{Error} &= \text{Actual Value} - \text{Predicted Value} \\ &= y - \hat{y}\end{aligned}$$

Graphically, the absolute value of the error is the vertical distance between the actual value of y (the point on the scatter diagram) and the predicted value of \hat{y} (the point on the linear regression line). In other words, the absolute value of the error measures the vertical distance between the actual data point and the line.



The slope and y -intercept for the linear regression equation are generated using the errors and the **least squares method**. The idea behind finding the line of best fit is based on the assumption that the data are scattered about a straight line. For any line, the errors can be calculated, squared, and then these squared errors can be added up. Of all of the possible lines, the line-of-best-fit is the **one** line that **minimizes** this sum of the squared errors. Any other line will have a higher sum of the squared errors compared to the sum of the squared errors for the line-of-best-fit.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=297#oembed-1>

Video: “Basic Excel Business Analytics #46: Slope & Intercept for Estimated Simple Linear Regression Equation” by excelisfun [18:29] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. What is the process through which we can calculate a line that goes through a scatter plot with a linear pattern?

Click to see Answer

Simple linear regression

2. An electronics retailer used regression to find a simple model to predict sales growth in the first quarter of the new year (January through March). The model is good for 90 days, where x is the day. The model can be written as $\hat{y} = 101.32 + 2.48x$ where \hat{y} is in thousands of dollars.
 - a. What would you predict the sales to be on day 60?
 - b. What would you predict the sales to be on day 90?

Click to see Answer

- a. \$250, 120
- b. \$342, 520

3. A landscaping company is hired to mow the grass for several large properties. The total area of the properties combined is 1,345 acres. The rate at which one person can mow is $\hat{y} = 1350 - 1.2x$ where x is the number of hours and \hat{y} represents the number of acres left to mow.
 - a. How many acres will be left to mow after 20 hours of work?
 - b. How many acres will be left to mow after 100 hours of work?
 - c. How many hours will it take to mow all of the lawns?

Click to see Answer

- a. 1,326
- b. 1,230
- c. 1,125

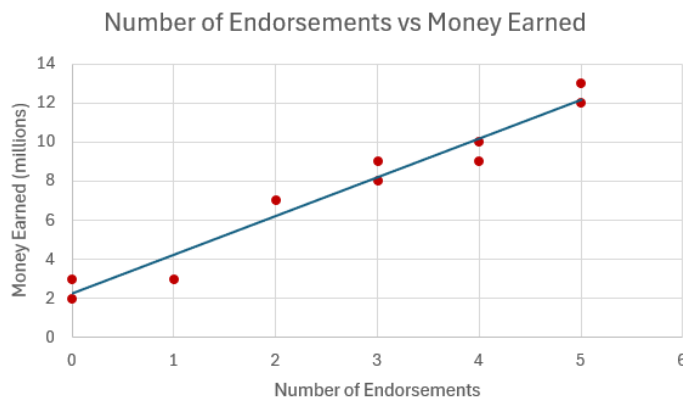
4. In a random sample of ten professional athletes, the number of endorsements the player has and the amount of money (in millions of dollars) the player earns are recorded in the table below.

Player	Number of Endorsements	Money Earned (in millions)
1	0	2
2	3	8
3	2	7
4	1	3
5	5	13
6	5	12
7	4	9
8	3	9
9	0	3
10	4	10

- Which variable is the independent variable, and which variable is the dependent variable?
- Use regression to find the equation for the line-of-best fit.
- Draw the scatter diagram for this data and include the line-of-best fit on the scatter diagram.
- What is the slope of the line of best fit? What does it represent?
- What is the y -intercept of the line-of-best-fit? What does it represent?
- Predict the amount of money a professional athlete earns if they have 2 endorsements.

Click to see Answer

- Independent: number of endorsements; Dependent: money earned
- $\hat{y} = 2.234 + 1.988x$ where x is the number of endorsements and \hat{y} is the money earned.



-
-
-
- 1.988. For each extra endorsement an athlete has, the amount of money earned

increases by \$1,988,000.

- e. 2.234. A player with 0 endorsements will earn \$2,234,000.
- f. \$6,208,723

5. The table below gives the percentage of workers who are paid hourly rates for the years 1979 to 1992. (Note: for identification of the independent and dependent variables, refer back to Question 7 in Section 12.2.)

Year	Percent of Workers Paid Hourly Rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

- a. Find the linear regression equation.
- b. Interpret the slope of the linear regression equation.
- c. What is the estimated percentage of workers paid hourly rates in 1988?

Click to see Answer

- a. $\hat{y} = -266.89 + 0.17x$ where x is the year and \hat{y} is the percent of workers paid an hourly rate.
- b. For each additional year, the percent of workers paid an hourly rate increases by 0.17%.
- c. 62.42%

6. The table below contains real data for the first two decades of AIDS cases. (Note: for identification of the independent and dependent variables, refer back to Question 1 in Section 12.2.)

Year	Number of AIDS Cases
1981	319
1982	1,170
1983	3,076
1984	6,240
1985	11,776
1986	19,032
1987	28,564
1988	35,447
1989	42,674
1990	48,634
1991	59,660
1992	78,530
1993	78,834
1994	71,874
1995	68,505
1996	59,347
1997	47,149
1998	38,393
1999	25,174
2000	25,522
2001	25,643
2002	26,464

- Find the linear regression equation.
- Interpret the slope of the linear regression equation.
- What is the predicted number of diagnosed cases for the year 1985?
- What is the predicted number of diagnosed cases for the year 1970? Why does this answer not make sense?

Click to see Answer

- $\hat{y} = -3,448,225.05 + 1749.78x$ where x is the year and \hat{y} is the number of AIDS cases.

- b. For each additional year, the number of AIDS cases increases by 1749.78.
 - c. 25,082.22
 - d. -1164.43. The number of AIDS cases is a count and so must be positive.
7. Recently, the annual number of driver deaths per 100,000 for the selected age groups was as shown in the table below. (Note: for identification of the independent and dependent variables, refer back to Question 8 in Section 12.2.)

Age	Number of Driver Deaths per 100,000
17.5	38
22	36
29.5	24
44.5	20
64.5	18
80	28

- a. Calculate the least squares (best-fit) line.
- b. Interpret the slope of the least squares line.
- c. Predict the number of driver deaths per 100,000 for people aged 40.

Click to see Answer

- a. $\hat{y} = 35.58 - 0.19x$ where x is the age and \hat{y} is the number of driver deaths per 100,000.
 - b. For each additional year of age, the number of driver deaths per 100,000 decreases by 0.19.
 - c. 27.91
8. The table below shows the life expectancy for an individual born in the United States in certain years. (Note: for identification of the independent and dependent variables, refer back to Question 9 in Section 12.2.)

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

- Find the linear regression equation.
- Interpret the slope of the linear regression equation.
- What is the estimated life expectancy for someone born in 1950? Why doesn't this value match the life expectancy given in the table for 1950?
- What is the estimated life expectancy for someone born in 1982?
- Using the regression equation, find the estimated life expectancy for someone born in 1850. Is this an accurate estimate for that year? Explain why or why not.

Click to see Answer

- $\hat{y} = -377.24 + 0.23x$ where x is the year and \hat{y} is life expectancy.
- For each additional year, the life expectancy increases by 0.23 years.
- 66.34. This is the value predicted by the model, which generally does not equal the actual value given in the data.
- 73.62 years
- 43.59 years. This is not an accurate estimate because the year 1850 is outside of the domain of the values of the independent variable provided in the data.

- The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level). (Note: for identification of the independent and dependent variables, refer back to Question 10 in Section 12.2.)

Height (in feet)	Number of Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

- Find the linear regression equation.
- Interpret the slope of the linear regression equation.
- What is the estimated height for a 32-story building?
- What is the estimated height for a 94-story building?
- Using the regression equation, find the estimated height for a 6-story building. Is this an accurate estimate for the height of a 6-story building? Explain why or why not.

Click to see Answer

- $\hat{y} = 102.43 + 11.76x$ where x is the number of stories and \hat{y} is the height.
- For each additional story, the height of the building increases by 11.76 feet.
- 478.70 feet
- 1207.73 feet
- 172.98 feet. This is not accurate because 6 is outside the domain of the independent variable given in the data.

- The following table shows data on average per capita wine consumption and heart disease rate in a random sample of 10 countries. (Note: for identification of the independent and dependent variables, refer back to Question 11 in Section 12.2.)

Per Capita Yearly Wine Consumption in Liters	Per Capita Death from Heart Disease
2.5	221
3.9	167
2.9	131
2.4	191
2.9	220
0.8	297
9.1	71
2.7	172
0.8	211
0.7	300

- Find the linear regression equation.
- Interpret the slope of the linear regression equation.
- What is the predicted per capita heart disease rate for a per capita yearly wine consumption of 2 litres?

Click to see Answer

- $\hat{y} = 266.63 - 23.88x$ where x is the per capita yearly wine consumption and \hat{y} is the per capita deaths from heart disease.
- For each additional litre of wine consumed per year, the number of deaths from heart disease decreases by 23.88.
- 218.87

- The following table consists of one student athlete's time (in minutes) to swim 2000 meters and the student's heart rate (beats per minute) after swimming on a random sample of 10 days. (Note: for identification of the independent and dependent variables, refer back to Question 12 in Section 12.2.)

Swim Time	Heart Rate
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

- Find the linear regression equation.
- Interpret the slope of the linear regression equation.
- What is the estimated heart rate for a swim time of 34.75 minutes?

Click to see Answer

- $\hat{y} = 193.88 - 1.49x$ where x is the swim time and \hat{y} is the heart rate.
- For each additional minute of swim time, the heart rate decreases by 1.49 beats per minute.
- 141.95 bpm

“12.5 The Regression Equation” and “12.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

12.5 COEFFICIENT OF DETERMINATION

LEARNING OBJECTIVES

- Calculate and interpret the coefficient of determination.

Previously, we saw how to use the correlation coefficient to measure the strength and direction of the linear relationship between the independent and dependent variables. The correlation coefficient gives us a way to measure how good a linear regression model fits the data. The coefficient of determination is another way to evaluate how well a linear regression model fits the data. Denoted r^2 , the **coefficient of determination** is the proportion of variation in the dependent variable that can be explained by the regression equation based on the independent variable. The coefficient of determination is the square of the correlation coefficient.

The coefficient of determination is a number between 0 and 1 and is the decimal form of a percent. The closer the coefficient of determination is to 1, the better the independent variable is at predicting the dependent variable. When we interpret the coefficient of determination, we use the percent form. When expressed as a percent, r^2 represents the percent of variation in the dependent variable y that can be explained by the variation in the independent variable x using the regression line. When interpreting the coefficient of determination, remember to be specific to the context of the question.

EXAMPLE

A statistics professor wants to study the relationship between a student's score on the third exam in the course and their final exam score. The professor took a random sample of 11 students and recorded their third exam score (out of 80) and their final exam score (out of 200). The results are recorded in the table below. The professor wants to develop a linear regression model to predict a student's final exam score from the third exam score.

Student	Third Exam Score	Final Exam Score
1	65	175
2	67	133
3	71	185
4	71	163
5	66	126
6	75	198
7	67	153
8	70	163
9	71	159
10	69	151
11	69	159

Previously we found the correlation coefficient $r = 0.6631$ and the line-of-best-fit $\hat{y} = -173.51 + 4.83x$ where x is the third exam score and \hat{y} is the (predicted) final exam score.

1. Find the coefficient of determination.
2. Interpret the coefficient of determination found in part 1.

Solution

1. $r^2 = (0.6631)^2 = 0.4397$.
2. 43.97% of the variation in the final exam score can be explained by the regression line based on the third exam score.

TRY IT

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the table below shows different depths with the maximum dive times in minutes. Previously, we found the correlation coefficient and the regression line to predict the maximum dive time from depth.

Depth (in feet)	Maximum Dive Time (in minutes)
50	80
60	55
70	45
80	35
90	25
100	22

1. Find the coefficient of determination.
2. Interpret the coefficient of determination found in part 1.

Click to see Solution

1. $r^2 = (-0.9629)^2 = 0.9272$.
2. 92.72% of the variation in the maximum dive time can be explained by the regression line based on depth.

Exercises

1. In a random sample of ten professional athletes, the number of endorsements the player has and the amount of money (in millions of dollars) the player earns are recorded in the table below. (Note: for identification of the independent and dependent variables, refer back to

Question 4 in Section 12.4.)

Player	Number of Endorsements	Money Earned (in millions)
1	0	2
2	3	8
3	2	7
4	1	3
5	5	13
6	5	12
7	4	9
8	3	9
9	0	3
10	4	10

- a. Calculate the coefficient of determination.
- b. Interpret the coefficient of determination.

Click to see Answer

- a. 0.9577
- b. 95.77\% of the money earned by an athlete can be explained by the regression line based on the number of endorsements.

2. The table below gives the percentage of workers who are paid hourly rates for the years 1979 to 1992. (Note: for identification of the independent and dependent variables, refer back to Question 7 in Section 12.2.)

Year	Percent of Workers Paid Hourly Rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

- a. Find the coefficient of determination.
- b. Interpret the coefficient of determination.

Click to see Answer

- a. 0.8926
- b. 89.26\% of the variation in the percent of workers paid an hourly rate can be explained by the regression line based on year.

3. The table below contains real data for the first two decades of AIDS cases. (Note: for identification of the independent and dependent variables, refer back to Question 1 in Section 12.2.)

Year	Number of AIDS Cases
1981	319
1982	1,170
1983	3,076
1984	6,240
1985	11,776
1986	19,032
1987	28,564
1988	35,447
1989	42,674
1990	48,634
1991	59,660
1992	78,530
1993	78,834
1994	71,874
1995	68,505
1996	59,347
1997	47,149
1998	38,393
1999	25,174
2000	25,522
2001	25,643
2002	26,464

- Calculate the coefficient of determination.
- Interpret the coefficient of determination.

Click to see Answer

- 0.2049
- 20.49\% of the variation in the number of AIDS cases is explained by the regression line based on the year.

4. Recently, the annual number of driver deaths per 100, 000 for the selected age groups was as shown in the table below. (Note: for identification of the independent and dependent variables, refer back to Question 8 in Section 12.2.)

Age	Number of Driver Deaths per 100, 000
17.5	38
22	36
29.5	24
44.5	20
64.5	18
80	28

- Find the coefficient of determination.
- Interpret the coefficient of determination.

Click to see Answer

- 0.3349
 - 33.49\% of the number of driver deaths per 100, 000 is explained by the regression line based on age.
5. The table below shows the life expectancy for an individual born in the United States in certain years. (Note: for identification of the independent and dependent variables, refer back to Question 9 in Section 12.2.)

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

- Calculate the coefficient of determination.
- Interpret the coefficient of determination.

Click to see Answer

- 0.9240
 - 92.40\% of the variation in life expectancy is explained by the regression line based on the year.
6. The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level). (Note: for identification of the independent and dependent variables, refer back to Question 10 in Section 12.2.)

Height (in feet)	Number of Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

- Calculate the coefficient of determination.
- Interpret the coefficient of determination.

Click to see Answer

- 0.8903
- 89.03\% of the variation in the height is explained by the regression line based on the number of stories.

- The following table shows data on average per capita wine consumption and heart disease rate in a random sample of 10 countries. (Note: for identification of the independent and dependent variables, refer back to Question 11 in Section 12.2.)

Per Capita Yearly Wine Consumption in Liters	Per Capita Death from Heart Disease
2.5	221
3.9	167
2.9	131
2.4	191
2.9	220
0.8	297
9.1	71
2.7	172
0.8	211
0.7	300

- Calculate the coefficient of determination.
- Interpret the coefficient of determination.

Click to see Answer

- 0.6987
- 69.87\% of the variation in the heart disease rate is explained by the regression line based on yearly wine consumption.

- The following table consists of one student athlete's time (in minutes) to swim 2000 meters and the student's heart rate (beats per minute) after swimming on a random sample of 10 days. (Note: for identification of the independent and dependent variables, refer back to Question 12 in Section 12.2.)

Swim Time	Heart Rate
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

- Calculate the coefficient of determination.
- Interpret the coefficient of determination.

Click to see Answer

- 0.0153
- 1.53\% of the variation in heart rate is explained by the regression line based on swim time.

“12.6 Coefficient of Determination” and “12.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

12.6 STANDARD ERROR OF THE ESTIMATE

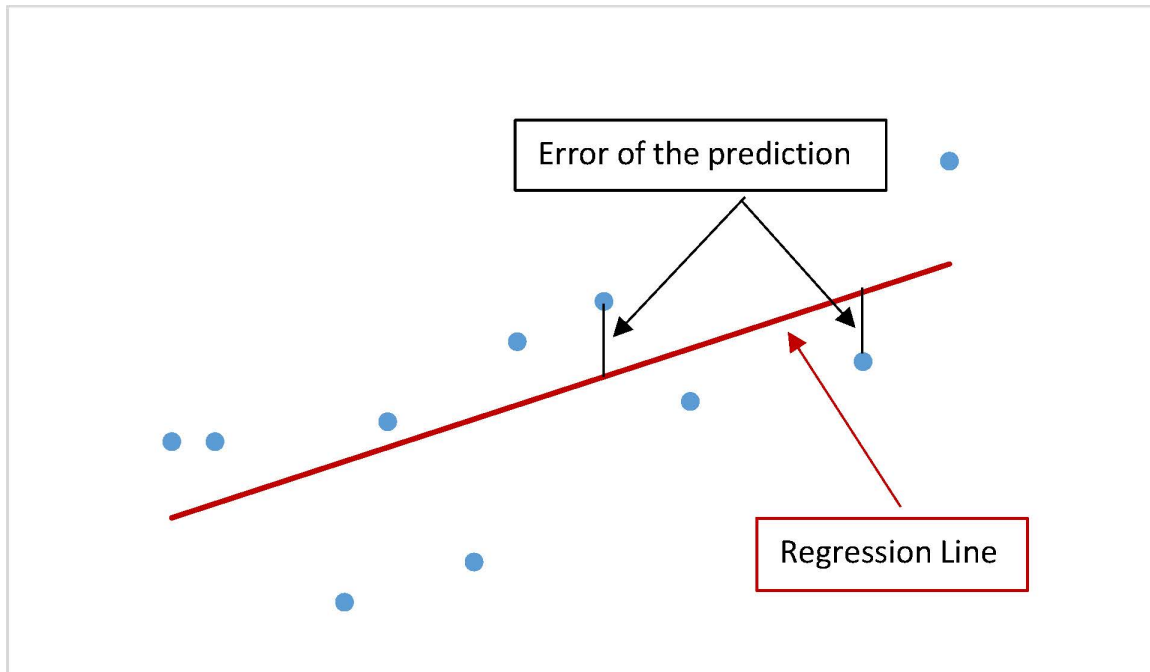
LEARNING OBJECTIVES

- Calculate and interpret the standard error of the estimate.

The difference between the actual value of the dependent variable y (in the sample data) and the predicted value of the dependent variable \hat{y} obtained from the linear regression equation is called the **error** or **residual**.

$$\text{Error} = \text{Actual Value} - \text{Predicted Value}$$

Graphically, the absolute value of the error is the vertical distance between the actual value of y (the point on the scatter diagram) and the predicted value of \hat{y} (the point on the linear regression line). In other words, the absolute value of the error measures the vertical distance between the actual data point and the line.



The **standard error of the estimate**, denoted s_e , is a measure of the standard deviation of the errors in a regression model. The standard error of the estimate is a measure of the average deviation or dispersion of the points on the scatter diagram around the line of best fit. The standard error of the estimate for the linear regression model is analogous to the standard deviation for a set of points, but instead of measuring the average distance from the mean, we are measuring the average distance from the regression line. Graphically, the standard error of the estimate measures the average vertical distance (the absolute value of the errors) between the points on the scatter diagram and the regression line.

When the points on the scatter diagram are close to the regression line, the errors are small, and so the average of the dispersion of the points around the line will be small. In this case, the value of the standard error of the estimate will be relatively small, which reflects the fact that there is little variation between the actual data points (the points on the scatter diagram) and the linear regression model. This implies that the linear regression model is a good fit for the data, and predictions made with the linear regression model will be fairly accurate.

Conversely, when the points on the scatter diagram are widely dispersed around the regression line, the errors are large, and so the average dispersion of the points around the line will be large. In this case, the value of the standard error of the estimate will be large, which reflects the greater dispersion between the actual data points and the linear regression model. This implies that the linear regression model is not a good fit for the data, and predictions made with the linear regression model will be inaccurate.

The value of s_e tells us, on average, how much the dependent variable differs from the regression line based on the independent variable. When interpreting the standard error of the estimate, remember to be specific to the question, using the actual names of the dependent and independent variables, and include appropriate units. The units of the standard error of the estimate are the same as the units of the dependent variable.

Although there is a formula to calculate the value of the standard error of the estimate, we will calculate the standard error of the estimate using the built-in function in Excel.

CALCULATING THE STANDARD ERROR OF THE ESTIMATE IN EXCEL

To calculate the standard error of the estimate, use the **steypx(array for y's, array for x's)** function.

- For **array for y's**, enter the cell array containing the **dependent** variable y data.
- For **array for x's**, enter the cell array containing the **independent** variable x data.

Visit the Microsoft page for more information about the **steypx** function.

NOTE

The order in which the data is entered into the **steypx** function is important. The data for the **dependent** variable is entered in the **first** array, and the data for the **independent** variable is entered in the **second** array. The output from the **steypx** function will be different when the order of the inputs is switched.

EXAMPLE

A statistics professor wants to study the relationship between a student's score on the third exam in the course and their final exam score. The professor took a random sample of 11 students and recorded their third exam score (out of 80) and their final exam score (out of 200). The results are recorded in the table below. The professor wants to develop a linear regression model to predict a student's final exam score from the third exam score.

Student	Third Exam Score	Final Exam Score
1	65	175
2	67	133
3	71	185
4	71	163
5	66	126
6	75	198
7	67	153
8	70	163
9	71	159
10	69	151
11	69	159

Previously, we found the line-of-best-fit $\hat{y} = -173.51 + 4.83x$ where x is the third exam score and \hat{y} is the (predicted) final exam score.

1. Find the standard error of the estimate.
2. Interpret the standard error of the estimate found in part 1.

Solution

1. Enter the data into an Excel spreadsheet. For this example, suppose we entered the data (without the column headings) so that the student column is in column A from A1 to A11, the third exam score is in column B from B1 to B11, and the final exam score is in column C from

C1 to C11.

Function	steypx
Field 1	C1:C11
Field 2	B1:B11
Answer	16.41

The value of the standard error of the estimate is $s_e = 16.41$.

- On average, the final exam score differs by **16.41** points from the regression line based on the third exam score.

TRY IT

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the table below shows different depths with the maximum dive times in minutes. Previously, we found the regression line to predict the maximum dive time from depth.

Depth (in feet)	Maximum Dive Time (in minutes)
50	80
60	55
70	45
80	35
90	25
100	22

- Find the standard error of the estimate.
- Interpret the standard error of the estimate found in part 1.

Click to see Solution

1. $s_e = 6.53$.
2. On average, the maximum dive time differs by **6.53** minutes from the regression line based on depth.

Exercises

1. In a random sample of ten professional athletes, the number of endorsements the player has and the amount of money (in millions of dollars) the player earns are recorded in the table below. (Note: for identification of the independent and dependent variables, refer back to Question 4 in Section 12.4.)

Player	Number of Endorsements	Money Earned (in millions)
1	0	2
2	3	8
3	2	7
4	1	3
5	5	13
6	5	12
7	4	9
8	3	9
9	0	3
10	4	10

- a. Calculate the standard error of the estimate.
- b. Interpret the standard error of the estimate.

Click to see Answer

- a. **0.8363**
- b. On average, the money earned by the athlete differs by **\$836, 300** from the regression

line based on the number of endorsements.

2. The table below gives the percentage of workers who are paid hourly rates for the years 1979 to 1992. (Note: for identification of the independent and dependent variables, refer back to Question 7 in Section 12.2.)

Year	Percent of Workers Paid Hourly Rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

- Find the standard error of the estimate.
- Interpret the standard error of the estimate.

Click to see Answer

- 0.2473
 - On average, the percent of workers paid an hourly rate differs by 0.2473% from the regression line based on year.
3. The table below contains real data for the first two decades of AIDS cases. (Note: for identification of the independent and dependent variables, refer back to Question 1 in Section 12.2.)

Year	Number of AIDS Cases
1981	319
1982	1,170
1983	3,076
1984	6,240
1985	11,776
1986	19,032
1987	28,564
1988	35,447
1989	42,674
1990	48,634
1991	59,660
1992	78,530
1993	78,834
1994	71,874
1995	68,505
1996	59,347
1997	47,149
1998	38,393
1999	25,174
2000	25,522
2001	25,643
2002	26,464

- Calculate the standard error of the estimate.
- Interpret the standard error of the estimate.

Click to see Answer

- 22, 936.69
- On average, the number of AIDS cases differs by 22, 936.69 from the regression line based on the year.

4. Recently, the annual number of driver deaths per 100, 000 for the selected age groups was as shown in the table below. (Note: for identification of the independent and dependent variables, refer back to Question 8 in Section 12.2.)

Age	Number of Driver Deaths per 100, 000
17.5	38
22	36
29.5	24
44.5	20
64.5	18
80	28

- Find the standard error of the estimate.
- Interpret the standard error of the estimate.

Click to see Answer

- 7.533
 - On average, the number of driver deaths per 100, 000 differs by 7.533 from the regression line based on age.
5. The table below shows the life expectancy for an individual born in the United States in certain years. (Note: for identification of the independent and dependent variables, refer back to Question 9 in Section 12.2.)

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

- a. Calculate the standard error of the estimate.
- b. Interpret the standard error of the estimate.

Click to see Answer

- a. 1.8202 years
- b. On average, the life expectancy differs by 1.8202 years from the regression line based on the year.

6. The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level). (Note: for identification of the independent and dependent variables, refer back to Question 10 in Section 12.2.)

Height (in feet)	Number of Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

- a. Calculate the standard error of the estimate.
- b. Interpret the standard error of the estimate

Click to see Answer

- a. 132.76 feet
- b. On average, the height differs by 132.76 feet from the regression line based on the number of stories.

7. The following table shows data on average per capita wine consumption and heart disease rate in a random sample of 10 countries. (Note: for identification of the independent and dependent variables, refer back to Question 11 in Section 12.2.)

Per Capita Yearly Wine Consumption in Liters	Per Capita Death from Heart Disease
2.5	221
3.9	167
2.9	131
2.4	191
2.9	220
0.8	297
9.1	71
2.7	172
0.8	211
0.7	300

- Calculate the standard error of the estimate.
- Interpret the standard error of the estimate.

Click to see Answer

- 40.563
 - On average, the number of deaths from heart disease differs by 40.563 from the regression line based on yearly wine consumption.
8. The following table consists of one student athlete's time (in minutes) to swim 2000 meters and the student's heart rate (beats per minute) after swimming on a random sample of 10 days. (Note: for identification of the independent and dependent variables, refer back to Question 12 in Section 12.2.)

Swim Time	Heart Rate
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

- Calculate the standard error of the estimate.
- Interpret the standard error of the estimate.

Click to see Answer

- 10.024 bpm
- On average, the heart rate differs by 10.024 bpm from the regression line based on swim time.

“12.7 Standard Error of the Estimate” and “12.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

PART XIII

MULTIPLE REGRESSION

Previously, we studied simple linear regression, which allowed us to build a model of the linear relationship between one independent variable and one dependent variable. Then, we could use the model to make predictions about the value of the dependent variable. For example, a simple linear regression model can be used to predict a person's salary (the dependent variable) from the person's age (the independent variable).

But, what if more than one independent variable impacts the value of the dependent variable? For example, a person's salary depends on more factors than just the person's age. A person's salary can also be related to their experience, their education, and their profession. We want to build a model that allows us to incorporate more than one independent variable. Because more information can be used in the model, additional independent variables can make regression models more accurate in predicting the dependent variable. A multiple regression model allows us to use two or more independent variables to predict one dependent variable.

As we saw with simple linear regression, in addition to building the model, we need ways to assess how good the multiple regression model fits the data and how good the model is at predicting values of the dependent variable.

CHAPTER OUTLINE

- 13.1 Multiple Regression
 - 13.2 Standard Error of the Estimate
 - 13.3 Coefficient of Multiple Determination
 - 13.4 Testing the Significance of the Overall Model
 - 13.5 Testing the Regression Coefficients
 - 13.6 Multicollinearity
-

“13.1 Introduction to Multiple Regression” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

13.1 MULTIPLE REGRESSION

LEARNING OBJECTIVES

- Develop a multiple regression model.
- Use a multiple regression model to predict values of the dependent variable.
- Interpret the partial regression coefficients.

Previously, we learned about simple linear regression, which models the linear relationship between one independent variable x and one dependent variable y . The equation for the regression line is:

$$\hat{y} = b_0 + b_1x$$

\hat{y} = predicted value of y

x = value of the independent variable

b_0 = y -intercept of the line

b_1 = slope of the line

Multiple regression is an extension of simple linear regression where there is still only one dependent variable y but two or more independent variables x_1, x_2, \dots, x_k . Multiple regression is motivated by scenarios where many independent variables may be simultaneously connected to a dependent variable. For example, the price of a product is related to the demand for the product, the time of year, and the competition's price.

The equation for the multiple regression model is:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

where \hat{y} is the predicted value of y , x_1, x_2, \dots, x_k are the independent variables, b_0, b_1, \dots, b_k are the regression coefficients, and k is the number of independent variables.

We will use Excel to generate the values of the regression coefficients. However, unlike simple linear regression, where we used individual, built-in functions to find the slope and y -intercept, we use a regression summary table to generate the values of the regression coefficients. As we will see, the regression summary table contains lots of information relating to the multiple regression model. For now, we will use the regression summary table to find the regression coefficients to create the multiple regression model. In later sections, we will learn about some of the other information contained in the regression summary table.

USING EXCEL TO CREATE A REGRESSION SUMMARY TABLE

In order to create a regression summary table, we need to use the Analysis ToolPak. Follow these instructions to add the Analysis ToolPak.

1. Enter the data into an Excel worksheet.
2. Go to the **Data** tab and click on **Data Analysis**. If you do not see **Data Analysis** in the **Data** tab, you will need to install the Analysis ToolPak.
3. In the **Data Analysis** window, select **Regression** and then click **OK**.
4. In the **Input Y Range**, enter the cell range for the y (dependent variable) data.
5. In the **Input X Range**, enter the cell range for the x (independent variables) data.
6. Click on **Labels in the first row** if you included the column headings in the input range.
7. From the **Output Options**, select the location where you want the output to appear. The default is a new worksheet.
8. Click **OK**. Excel will then generate a regression summary table.

NOTES

1. For the **Input X Range**, the data for the independent variables must all be together. That is,

the columns (or rows) containing the data for the independent variables must all be consecutive. If the column (or row) containing data for the dependent variable is in between two columns (or rows) containing independent variables, copy the dependent variable column and paste the dependent variable column at the beginning or end of the columns (or rows) of data. Make sure to delete the original dependent variable column/row after placing a copy at the beginning or end of the data.

2. There are several other options available in the Regression input window, such as for confidence intervals or information about residuals. We will not need any of this other information, so leave everything else unchecked.
3. This website provides a detailed explanation of the information contained in the regression summary table.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken, and the data is recorded in the table below. The employee's income is recorded in \$1000s, and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction. Develop a multiple regression model to predict the job satisfaction score from the other variables.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Solution

There are three independent variables: hours of unpaid work per week, age, and income (\$1000s).

Let x_1 be the hours of unpaid work per week, let x_2 be age, and let x_3 be income (\$1000s). The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The coefficients for the multiple regression model are in the **Coefficients** column in the bottom part of the table. The value of b_0 is in the Intercept row, so $b_0 = 4.7993$. The value of b_1 , the coefficient for x_1 , is in the Hours of Unpaid Work per Week row, so $b_1 = -0.3818$. The value of b_2 , the coefficient of x_2 , is in the Age row, so $b_2 = 0.0046$. The value of b_3 , the coefficient of x_3 , is in the Income row, so $b_3 = 0.0233$.

The multiple regression equation is

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

NOTES

1. When writing down the multiple regression equation, remember to define what the variables represent in the context of the question. That is, state what \hat{y} and the independent variables represent in relation to the question.
2. A couple of the columns on the right side of the regression summary table generated by Excel were deleted in order to fit the table onto the page. These columns are not necessary for the work we will be doing.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=311#oembed-1>

Video: “Basic Excel Business Analytics #50: Introduction to Multiple Regression, Data Analysis Regression” by excelisfun [13:34] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Regression Coefficients

Recall that the slope b_1 in the simple linear regression model $\hat{y} = b_0 + b_1x$ tells us how the dependent variable y changes for a single unit increase in the independent variable x . In a similar way, each regression coefficient b_i represents the change (increase or decrease) in the dependent variable for a one-unit increase in the corresponding independent variable x_i , while all the other variables are held constant. When interpreting a regression coefficient, it is important to be specific to the question, using the actual names of the variables and correct units.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken, and the data is recorded in the table below. The employee's income is recorded in \$1000s, and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60

7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

1. Interpret the regression coefficient for hours of unpaid work per week.
2. Interpret the regression coefficient for age.
3. Interpret the regression coefficient for income.

Solution

1. $b_1 = -0.3818$. Interpretation: For a one-hour increase in the hours of unpaid work per week, the job satisfaction score decreases by 0.3818, while the other variables are held constant.
2. $b_2 = 0.0046$. Interpretation: For a one-year increase in the age of the employee, the job satisfaction score increases by 0.0046, while the other variables are held constant.
3. $b_3 = 0.0233$. Interpretation: For a \$1000 increase in income, the job satisfaction score increases by 0.0233, while the other variables are held constant.

NOTES

1. Remember to include “while the other variables are held constant” with the interpretation of each regression coefficient. We can only talk about how the change in one independent variable affects the dependent variable, so the values of the other variables must be kept fixed.

2. When writing down the interpretation of each regression coefficient, remember to be specific to the question using the actual names of the independent and dependent variables and appropriate units.
3. Each regression coefficient has the same units as the dependent variable.
4. Income is measured in \$1000s, so a one-unit increase in the income variable actually corresponds to a \$1000 increase in income.

Making Predictions with a Multiple Regression Model

As with simple linear regression, a multiple regression model can be used to make predictions about the dependent variable from specific values of the independent variables. To make a prediction, substitute the corresponding values of the independent variables into the multiple regression equation and calculate out the value of \hat{y} . Watch out for the units of measurement for each variable when using the multiple regression equation—the units of the values entered into the independent variable x_i in the multiple regression equation must match the units of the independent variable in the sample data.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken, and the data is recorded in the table below. The employee's income is recorded in \$1000s, and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60

7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

Predict the job satisfaction score for a 40-year-old employee who works two hours of unpaid work per week and has an income of \$75,000.

Solution

The values of the independent variables we need to enter into the multiple regression model are

$x_1 = 2$, $x_2 = 40$, and $x_3 = 75$:

$$\begin{aligned}\hat{y} &= 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3 \\ &= 4.7993 - 0.3818 \times 2 + 0.0046 \times 40 + 0.0233 \times 75 \\ &= 5.96\end{aligned}$$

The predicted job satisfaction score for a 40-year-old employee who works two hours of unpaid work per week and has an income of \$75,000 is 5.96.

NOTES

1. In the sample data, income is measured in \$1000s. So an income of \$75,000 would be recorded as 75 in the sample data. So, we enter 75 for the value of x_3
2. To get the most accurate answer, use Excel to calculate out the value of \hat{y} , clicking on the corresponding cells containing the values of the coefficients in the regression summary sheet.

Assumptions about the Multiple Regression Model

The multiple regression model given above is the model we create from **sample data**—a sample is taken from the population, and the sample data is used to find the regression coefficients in the model. So the regression coefficients, b_0, b_1, \dots, b_k , are estimates of the corresponding population parameters for the regression coefficients, $\beta_0, \beta_1, \dots, \beta_k$.

The population model for the multiple regression equation is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

where x_1, x_2, \dots, x_k are the independent variables, $\beta_0, \beta_1, \dots, \beta_k$ are the population parameters of the regression coefficients, and ϵ is the error variable. Because the independent variables do not account for all of the variability in the dependent variable y , the error variable ϵ captures the effects of variables other than the independent variables.

We must make certain assumptions about the regression model, in particular about the errors/residuals for the population data, in order to obtain valid conclusions about the multiple regression model. Recall that the error or residual is the difference between the actual value of y in the sample data and the predicted value \hat{y} from the model. Because we do not have the population data to work with, we cannot verify if these conditions are met. However, much of regression analysis, including testing how well the data fit the model, depends on these assumptions being true.

Assumptions about the multiple regression model include:

- The model is linear.
- The errors/residuals have a normal distribution.
- The mean of the errors/residuals is 0.
- The variance of the errors/residuals is constant.
- The errors/residuals are independent.

Exercises

1. A local restaurant advocacy group wants to study the relationship between a restaurant's average weekly profit, the restaurant's seating capacity, and the average daily traffic that passes the restaurant's location. The group took a sample of restaurants and recorded their

average weekly profit (in \$1000s), the seating restaurant's seating capacity, and the average number of cars (in 1000s) that passes the restaurant's location. The data is recorded in the following table:

Seating Capacity	Traffic Count (1000s)	Weekly Net Profit (\$1000s)
120	19	23.8
180	8	29.2
150	12	22
180	15	26.2
220	16	33.5
235	10	32
115	18	22.4
110	12	20.4
165	21	23.7
220	20	34.7
140	24	27.1
145	24	23.3
140	13	20.9
200	14	29.6
210	14	31.4
175	12	23.2
175	15	31.1
190	17	28.2
100	23	25.2
145	20	20.7
135	13	37.2
25	13	26.3
140	25	20
130	14	28.2
135	10	24.6
160	23	23.7

- Find the regression model to predict the average weekly profit from the other variables.
- Interpret the coefficient for seating capacity.

- c. Interpret the coefficient for traffic count.
- d. Predict the average weekly profit for a restaurant with a seating capacity of 150 and a traffic count of 25,000 cars.

Click to see Answer

$$\hat{y} = 21.989 + 0.046x_1 - 0.196x_2$$

- a. x_1 = seating capacity
 x_2 = traffic count (1000s)
 \hat{y} = average weekly profit (\$1000s)
- b. For each additional seat in the restaurant, the average weekly profit increases by \$46.
- c. For each additional 1000 cars that pass the restaurant, the average weekly profit decreases by \$196.
- d. \$24,519.20

2. A local university wants to study the relationship between a student's GPA, the average number of hours they spend studying each night, and the average number of nights they go out each week. The university took a sample of students and recorded the following data:

GPA	Average Number of Hours Spent Studying Each Night	Average Number of Nights Go Out Each Week
3.72	5	1
3.88	3	1
3.67	2	1
3.87	3	4
2.49	1	4
1.29	1	2
1.01	2	4
2.12	1	1
1.9	1	5
3.42	3	2
1.33	1	4
1.07	0	2
2.75	3	1
3.82	4	1
3.91	5	0
2.25	2	3
2.06	1	5
2.92	3	2
3.06	3	1
3.65	2	2
3.69	4	1

- Find the regression model to predict GPA from the other variables.
- Interpret the coefficient for the average number of hours spent studying each night.
- Interpret the coefficient for the average number of nights a student goes out each week.
- Predict the GPA for a student who spends an average of 4 hours a night studying and goes out an average of 3 nights a week.

Click to see Answer

$$\hat{y} = 1.692 + 0.524x_1 - 0.082x_2$$

a. x_1 = average number of hours spent studying a night

x_2 = average number of nights go out each week

$$\hat{y} = \text{GPA}$$

b. For each additional hour spent studying each night, the student's GPA increases by 0.524.

c. For each additional hour a student goes out each week, the student's GPA decreases by 0.082.

d. 3.54

3. A very large company wants to study the relationship between the salaries of employees in management positions, their age, the number of years the employee spent in college, and the number of years the employee has been with the company. A sample of management employees is taken, and the data is recorded below:

Age	Years of College	Years with Company	Salary (\$1000s)
60	8	29	317.3
33	3	5	97.3
57	6	27	263.1
32	4	5	101.3
31	6	3	114.2
61	8	19	350.4
41	7	8	146.9
35	4	2	91.7
51	6	21	198.2
50	8	10	196.5
57	5	15	105.7
49	6	18	118.3
62	7	27	305.2
52	8	26	239.9
39	4	8	145.9
42	7	5	175.4
62	4	24	219.4
60	4	22	202.1
65	3	21	196.3
40	4	10	143.9
62	6	29	408.7
53	7	5	145.2
48	8	5	175.1
61	5	6	152.7
38	7	3	99.7

40	7	12	174.9
45	7	7	149.2
58	7	14	282.8
38	4	3	95.7
41	5	18	232.8

- Find the regression model to predict salary from the other variables.
- Interpret the coefficient for age.
- Interpret the coefficient for years of college.
- Interpret the coefficient for years with the company.
- Predict the salary for a 47-year-old management employee who spent 5 years in college and has been with the company for 15 years.

Click to see Answer

$$\hat{y} = -42.359 + 1.436x_1 + 14.758x_2 + 5.486x_3$$

x_1 = age

- x_2 = years of college

x_3 = years with the company

\hat{y} = salary (\$1000s)

- For each additional year of age, the salary increases by \$1436.14.
- For each additional year of college, the salary increases by \$14,758.04.
- For each additional year with the company, the salary increases by \$5486.07.
- \$181,221.15

“13.2 Multiple Regression” and “13.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

13.2 STANDARD ERROR OF THE ESTIMATE

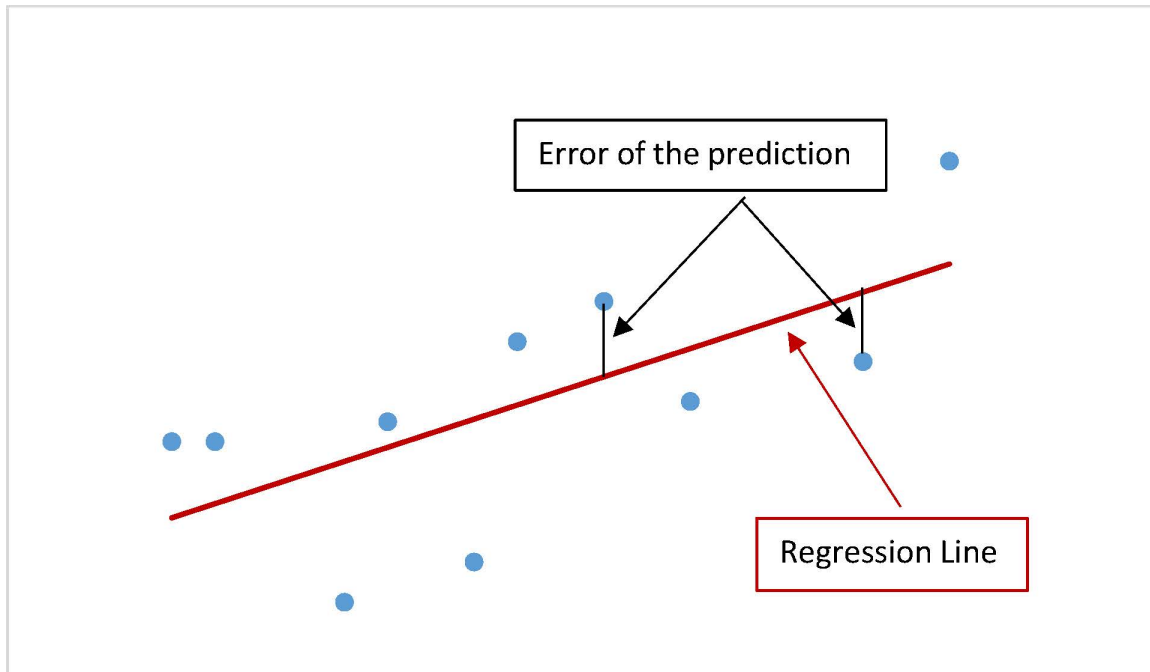
LEARNING OBJECTIVES

- Calculate and interpret the standard error of the estimate for multiple regression.

The difference between the actual value of the dependent variable y (in the sample data) and the predicted value of the dependent variable \hat{y} obtained from the multiple regression model is called the **error** or **residual**.

$$\text{Error} = \text{Actual Value} - \text{Predicted Value}$$

For the simple linear regression model, the standard error of the estimate measures the average vertical distance (the error) between the points on the scatter diagram and the regression line.



The **standard error of the estimate**, denoted s_e , is a measure of the standard deviation of the errors in a regression model. The standard error of the estimate is a measure of the average deviation of the errors, the difference between the \hat{y} -values predicted by the multiple regression model and the y -values in the sample. The standard error of the estimate for the regression model is the standard deviation of the errors/residuals.

The value of s_e tells us, on average, how much the dependent variable differs from the regression model based on the independent variables. When interpreting the standard error of the estimate, remember to be specific to the question, using the actual names of the dependent and independent variables, and include appropriate units. The units of the standard error of the estimate are the same as the units of the dependent variable.

The value of the standard error of the estimate for the regression model can be found in the regression summary table, which we learned how to generate in Excel in the previous section.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken, and the data is recorded in the table below. The employee's income is recorded in \$1000s, and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60

7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

1. Find the standard error of the estimate.
2. Interpret the standard error of the estimate.

Solution

1. The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The standard error of the estimate for the regression model is in the top part of the table, under the **Regression Statistics** heading in the **Standard Error** row. The value of the standard error of the estimate is $s_e = 1.5852$.

- On average, the job satisfaction score is 1.5852 points away from the regression model based on the independent variables “hours of unpaid work per week,” “age,” and “income.”

NOTE

The standard error of the estimate for the regression model is located in the **top** part of the table under the **Regression Statistics** heading. You will notice another standard error column at the bottom in the rows corresponding to the independent variables. These standard errors in the bottom part of the table are not related to the standard error of the estimate. In fact, the standard errors in the independent variable rows are measures of the uncertainty around the estimate of the regression coefficient for each independent variable.

Exercises

1. A local restaurant advocacy group wants to study the relationship between a restaurant's average weekly profit, the restaurant's seating capacity, and the average daily traffic that passes the restaurant's location. The group took a sample of restaurants and recorded their average weekly profit (in \$1000s), the seating restaurant's seating capacity, and the average number of cars (in 1000s) that passes the restaurant's location. The data is recorded in the following table:

Seating Capacity	Traffic Count (1000s)	Weekly Net Profit (\$1000s)
120	19	23.8
180	8	29.2
150	12	22
180	15	26.2
220	16	33.5
235	10	32
115	18	22.4
110	12	20.4
165	21	23.7
220	20	34.7
140	24	27.1
145	24	23.3
140	13	20.9
200	14	29.6
210	14	31.4
175	12	23.2
175	15	31.1
190	17	28.2
100	23	25.2
145	20	20.7
135	13	37.2
25	13	26.3
140	25	20
130	14	28.2
135	10	24.6
160	23	23.7

In Question 1 of Section 13.1, we found the regression model to predict the average weekly profit from other variables.

- a. Find the standard error of the estimate for the regression model.
- b. Interpret the standard error of the estimate.

Click to see Answer

- a. 4.1675
 - b. On average, the average weekly profit differs by \$4, 167.50 from the regression model based on seating capacity and traffic count.
-
2. A local university wants to study the relationship between a student's GPA, the average number of hours they spend studying each night, and the average number of nights they go out each week. The university took a sample of students and recorded the following data:

GPA	Average Number of Hours Spent Studying Each Night	Average Number of Nights Go Out Each Week
3.72	5	1
3.88	3	1
3.67	2	1
3.87	3	4
2.49	1	4
1.29	1	2
1.01	2	4
2.12	1	1
1.9	1	5
3.42	3	2
1.33	1	4
1.07	0	2
2.75	3	1
3.82	4	1
3.91	5	0
2.25	2	3
2.06	1	5
2.92	3	2
3.06	3	1
3.65	2	2
3.69	4	1

In Question 2 of Section 13.1, we found the regression model to predict GPA from other variables.

- Find the standard error of the estimate for the regression model.
- Interpret the standard error of the estimate.

Click to see Answer

- 0.6613
- On average, GPA differs by 0.6613 from the regression model based on the average number of hours spent studying a night and the average number of nights a student goes

out each week.

3. A very large company wants to study the relationship between the salaries of employees in management positions, their age, the number of years the employee spent in college, and the number of years the employee has been with the company. A sample of management employees is taken, and the data is recorded below:

Age	Years of College	Years with Company	Salary (\$1000s)
60	8	29	317.3
33	3	5	97.3
57	6	27	263.1
32	4	5	101.3
31	6	3	114.2
61	8	19	350.4
41	7	8	146.9
35	4	2	91.7
51	6	21	198.2
50	8	10	196.5
57	5	15	105.7
49	6	18	118.3
62	7	27	305.2
52	8	26	239.9
39	4	8	145.9
42	7	5	175.4
62	4	24	219.4
60	4	22	202.1
65	3	21	196.3
40	4	10	143.9
62	6	29	408.7
53	7	5	145.2
48	8	5	175.1
61	5	6	152.7
38	7	3	99.7

40	7	12	174.9
45	7	7	149.2
58	7	14	282.8
38	4	3	95.7
41	5	18	232.8

In Question 3 of Section 13.1, we found the regression model to predict salary from other variables.

- Find the standard error of the estimate for the regression model.
- Interpret the standard error of the estimate.

Click to see Answer

- 45.24522
- On average, salary differs by \$45, 255.22 from the regression model based on age, years of college, and years with the company.

“13.3 Standard Error of the Estimate” and “13.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

13.3 COEFFICIENT OF MULTIPLE DETERMINATION

LEARNING OBJECTIVES

- Calculate and interpret the coefficient of multiple determination.
- Calculate and interpret the adjusted coefficient of multiple determination.

Previously, we learned about the coefficient of determination, r^2 , for simple linear regression, which is the proportion of variation in the dependent variable that can be explained by the simple linear regression model based on the independent variable. The coefficient of determination is a good way to measure how well the simple linear regression model fits the data.

Coefficient of Multiple Determination

The **coefficient of multiple determination**, denoted R^2 , in multiple regression is similar to the coefficient of determination in simple linear regression, except in multiple regression, there is more than one independent variable. The coefficient of multiple determination is the proportion of variation in the dependent variable that can be explained by the multiple regression model based on the independent variables.

The value of the coefficient of multiple determination is found on the regression summary table, which we learned how to generate in Excel in a previous section. We interpret the coefficient of multiple determination in the same way that we interpret the coefficient of determination for simple linear regression.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken, and the data is recorded in the table below. The employee's income is recorded in \$1000s, and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

1. Find the coefficient of multiple determination.
2. Interpret the coefficient of multiple determination.

Solution

1. The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The coefficient of multiple determination for the regression model is in the top part of the table, under the **Regression Statistics** heading in the **R Square** row. The value of the coefficient of multiple determination is $R^2 = 0.5066$.

- 50.66\% of the variation in the job satisfaction score can be explained by the regression model based on the independent variables “hours of unpaid work per week,” “age,” and “income.”

Adjusted Coefficient of Multiple Determination

The value of the coefficient of multiple determination **always** increases as more independent

variables are added to the model, even if the new independent variable has no relationship with the dependent variable. The coefficient of multiple determination is an inflated value when additional independent variables do not add any significant information to the dependent variable. Consequently, the coefficient of multiple determination is an **overestimate** of the contribution of the independent variables when new independent variables are added to the model.

Instead, we use the **adjusted coefficient of multiple determination**, denoted $\text{adjusted } R^2$, which corrects the overestimation of the coefficient of multiple determination when new independent variables are added to the model. The adjusted coefficient of multiple determination is interpreted in the same way as the coefficient of multiple determination. The adjusted coefficient of multiple determination adjusts the value of R^2 to account for the number of independent variables in the model in order to avoid overestimating the impact of adding independent variables to the model.

The adjusted coefficient of multiple determination is calculated from the value of R^2 :

$$\text{adjusted } R^2 = 1 - \left(\frac{(n - 1) \times (1 - R^2)}{n - k - 1} \right)$$

where n is the number of observations and k is the number of independent variables. Although we can find the value of the adjusted coefficient of multiple determination using the above formula, the value of the coefficient of multiple determination is found on the regression summary table.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken, and the data is recorded in the table below. The employee's income is recorded in \$1000s, and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

1. Find the adjusted coefficient of multiple determination.
2. Interpret the adjusted coefficient of multiple determination.

Solution

1. The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The adjusted coefficient of multiple determination for the regression model is in the top part of the table, under the **Regression Statistics** heading in the **Adjusted R Square** row. The value of the adjusted coefficient of multiple determination is adjusted $R^2 = 0.4361$.

- 43.61\% of the variation in the job satisfaction score can be explained by the regression model based on the independent variables “hours of unpaid work per week,” “age,” and “income.”

If the addition of a new independent variable increases the value of the adjusted coefficient of multiple determination, then this is an indication that the regression model has improved as

a result of adding the new independent variable. But, if the addition of a new independent variable decreases the value of the adjusted coefficient of multiple determination, then the added independent variable has not improved the overall regression model. In such cases, the new independent variable should not be added to the model.

Exercises

1. A local restaurant advocacy group wants to study the relationship between a restaurant's average weekly profit, the restaurant's seating capacity, and the average daily traffic that passes the restaurant's location. The group took a sample of restaurants and recorded their average weekly profit (in \$1000s), the seating restaurant's seating capacity, and the average number of cars (in 1000s) that passes the restaurant's location. The data is recorded in the following table:

Seating Capacity	Traffic Count (1000s)	Weekly Net Profit (\$1000s)
120	19	23.8
180	8	29.2
150	12	22
180	15	26.2
220	16	33.5
235	10	32
115	18	22.4
110	12	20.4
165	21	23.7
220	20	34.7
140	24	27.1
145	24	23.3
140	13	20.9
200	14	29.6
210	14	31.4
175	12	23.2
175	15	31.1
190	17	28.2
100	23	25.2
145	20	20.7
135	13	37.2
25	13	26.3
140	25	20
130	14	28.2
135	10	24.6
160	23	23.7

In Question 1 of Section 13.1, we found the regression model to predict the average weekly profit from other variables.

- a. Find the adjusted coefficient of determination for the regression model.
- b. Interpret the adjusted coefficient of determination.

Click to see Answer

- a. 0.2250
 - b. 22.50\% of the variation in the average weekly profit can be explained by the regression model based on seating capacity and traffic count.
-
2. A local university wants to study the relationship between a student's GPA, the average number of hours they spend studying each night, and the average number of nights they go out each week. The university took a sample of students and recorded the following data:

GPA	Average Number of Hours Spent Studying Each Night	Average Number of Nights Go Out Each Week
3.72	5	1
3.88	3	1
3.67	2	1
3.87	3	4
2.49	1	4
1.29	1	2
1.01	2	4
2.12	1	1
1.9	1	5
3.42	3	2
1.33	1	4
1.07	0	2
2.75	3	1
3.82	4	1
3.91	5	0
2.25	2	3
2.06	1	5
2.92	3	2
3.06	3	1
3.65	2	2
3.69	4	1

In Question 2 of Section 13.1, we found the regression model to predict GPA from other variables.

- Find the adjusted coefficient of determination for the regression model.
- Interpret the adjusted coefficient of determination for the regression model.

Click to see Answer

- 0.5833
- 58.33\% of the variation in GPA can be explained by the regression model based on the average number of hours spent studying a night and the average number of nights a

student goes out each week.

3. A very large company wants to study the relationship between the salaries of employees in management positions, their age, the number of years the employee spent in college, and the number of years the employee has been with the company. A sample of management employees is taken, and the data is recorded below:

Age	Years of College	Years with Company	Salary (\$1000s)
60	8	29	317.3
33	3	5	97.3
57	6	27	263.1
32	4	5	101.3
31	6	3	114.2
61	8	19	350.4
41	7	8	146.9
35	4	2	91.7
51	6	21	198.2
50	8	10	196.5
57	5	15	105.7
49	6	18	118.3
62	7	27	305.2
52	8	26	239.9
39	4	8	145.9
42	7	5	175.4
62	4	24	219.4
60	4	22	202.1
65	3	21	196.3
40	4	10	143.9
62	6	29	408.7
53	7	5	145.2
48	8	5	175.1
61	5	6	152.7
38	7	3	99.7

40	7	12	174.9
45	7	7	149.2
58	7	14	282.8
38	4	3	95.7
41	5	18	232.8

In Question 3 of Section 13.1, we found the regression model to predict salary from other variables.

- Find the adjusted coefficient of determination for the regression model.
- Interpret the adjusted coefficient of determination.

Click to see Answer

- 0.6959
- 69.59\% of the variation in salary can be explained by the regression model based on age, years of college, and years with the company.

“13.4 Coefficient of Multiple Determination” and “13.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

13.4 TESTING THE SIGNIFICANCE OF THE OVERALL MODEL

LEARNING OBJECTIVES

- Conduct and interpret an overall model test on a multiple regression model.

Previously, we learned that the population model for the multiple regression equation is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

where x_1, x_2, \dots, x_k are the independent variables, $\beta_0, \beta_1, \dots, \beta_k$ are the population parameters of the regression coefficients, and ϵ is the error variable. The error variable ϵ accounts for the variability in the dependent variable that is not captured by the linear relationship between the dependent and independent variables. The value of ϵ cannot be determined, but we must make certain assumptions about ϵ and the errors/residuals in the model in order to conduct a hypothesis test on how well the model fits the data. These assumptions include:

- The model is linear.
- The errors/residuals have a normal distribution.
- The mean of the errors/residuals is 0.
- The variance of the errors/residuals is constant.
- The errors/residuals are independent.

Because we do not have the population data, we cannot verify that these conditions are met. We need to assume that the regression model has these properties in order to conduct hypothesis tests on the model.

Testing the Overall Model

We want to test if there is a relationship between the dependent variable and the **set** of independent variables. In other words, we want to determine if the regression model is valid or invalid.

- **Invalid Model.** There is no relationship between the dependent variable and the set of independent variables. In this case, **all** of the regression coefficients β_i in the population model are zero. This is the claim for the null hypothesis in the overall model test:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

- **Valid Model.** There is a relationship between the dependent variable and the set of independent variables. In this case, **at least one** of the regression coefficients β_i in the population model is not zero. This is the claim for the alternative hypothesis in the overall model test: $H_a : \text{at least one } \beta_i \neq 0.$

The overall model test procedure compares the means of explained and unexplained variation in the model in order to determine if the explained variation (caused by the relationship between the dependent variable and the set of independent variables) in the model is larger than the unexplained variation (represented by the error variable ϵ). If the explained variation is larger than the unexplained variation, then there is a relationship between the dependent variable and the set of independent variables, and the model is valid. Otherwise, there is no relationship between the dependent variable and the set of independent variables, and the model is invalid.

The logic behind the overall model test is based on two independent estimates of the variance of the errors:

- One estimate of the variance of the errors, MSR , is based on the mean amount of explained variation in the dependent variable y .
- One estimate of the variance of the errors, MSE , is based on the mean amount of unexplained variation in the dependent variable y .

The overall model test compares these two estimates of the variance of the errors to determine if there is a relationship between the dependent variable and the set of independent variables. Because the overall model test involves the comparison of two estimates of variance, an F -distribution is used to conduct the overall model test, where the test statistic is the ratio of the two estimates of the variance of the errors.

The **mean square due to regression**, MSR , is one of the estimates of the variance of the errors. The MSR is the estimate of the variance of the errors determined by the variance of the predicted

\hat{y} -values from the regression model and the mean of the y -values in the sample, \bar{y} . If there is no relationship between the dependent variable and the set of independent variables, then the MSR provides an unbiased estimate of the variance of the errors. If there is a relationship between the dependent variable and the set of independent variables, then the MSR provides an overestimate of the variance of the errors.

$$SSR = \sum (\hat{y} - \bar{y})^2$$

$$MSR = \frac{SSR}{k}$$

The **mean square due to error**, MSE , is the other estimate of the variance of the errors. The MSE is the estimate of the variance of the errors determined by the error $(y - \hat{y})$ in using the regression model to predict the values of the dependent variable in the sample. The MSE always provides an unbiased estimate of the variance of errors, regardless of whether or not there is a relationship between the dependent variable and the set of independent variables.

$$SSE = \sum (y - \hat{y})^2$$

$$MSE = \frac{SSE}{n - k - 1}$$

The overall model test depends on the fact that the MSR is influenced by the explained variation in the dependent variable, which results in the MSR being either an unbiased or overestimate of the variance of the errors. Because the MSE is based on the unexplained variation in the dependent variable, the MSE is not affected by the relationship between the dependent variable and the set of independent variables and is always an unbiased estimate of the variance of the errors.

The null hypothesis in the overall model test is that there is no relationship between the dependent variable and the set of independent variables. The alternative hypothesis is that there is a relationship between the dependent variable and the set of independent variables. The F -score for the overall model test is the ratio of the two estimates of the variance of the errors, $F = \frac{MSR}{MSE}$ with $df_1 = k$ and $df_2 = n - k - 1$. The p -value for the test is the area in the right tail of the F -distribution to the right of the F -score.

NOTES

1. If there is no relationship between the dependent variable and the set of independent variables, both the MSR and the MSE are unbiased estimates of the variance of the errors. In this case, the MSR and the MSE are close in value, which results in an F -score close to 1 and a large p -value. The conclusion of the test would be that the null hypothesis is true.
2. If there is a relationship between the dependent variable and the set of independent variables, the MSR is an overestimate of the variance of the errors. In this case, the MSR is significantly larger than the MSE , which results in a large F -score and a small p -value. The conclusion of the test would be that the alternative hypothesis is true.

Conducting a Hypothesis Test on the Overall Regression Model

Follow these steps to perform a hypothesis test on the overall regression model:

1. Write down the null hypothesis that there is no relationship between the dependent variable and the set of independent variables:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

2. Write down the alternative hypotheses that there is a relationship between the dependent variable and the set of independent variables:

$$H_a : \text{at least one } \beta_i \text{ is not } 0$$

3. Collect the sample information for the test and identify the significance level α .
4. The p -value is the area in the right tail of the F -distribution. The F -score and degrees of freedom are

$$F = \frac{MSR}{MSE}$$

$$df_1 = k$$

$$df_2 = n - k - 1$$

5. Compare the p – value to the significance level and state the outcome of the test.
 - If p – value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If p – value $> \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
6. Write down a concluding sentence specific to the context of the question.

The calculation of the MSR , the MSE , and the F -score for the overall model test can be time-consuming, even with the help of software like Excel. However, the required F -score and p – value for the test can be found on the regression summary table, which we learned how to generate in Excel in a previous section.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken, and the data is recorded in the table below. The employee's income is recorded in \$1000s, and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60

7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

At the 5% significance level, test the validity of the overall model to predict the job satisfaction score.

Solution

Hypotheses:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_a : \text{at least one } \beta_i \text{ is not } 0$$

p – value:

The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The p – value for the overall model test is in the middle part of the table under the **ANOVA** heading in the **Significance F column** of the **Regression row**. So the p – value = 0.0017.

Conclusion:

Because p – value = $0.0017 < 0.05 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 5\% significance level, there is enough evidence to suggest that there is a relationship between the dependent variable “job satisfaction” and the set of independent variables “hours of unpaid work per week,” “age”, and “income.”

NOTES

1. The null hypothesis $\beta_1 = \beta_2 = \beta_3 = 0$ is the claim that all of the regression coefficients are zero. That is, the null hypothesis is the claim that there is **no** relationship between the dependent variable and the set of independent variables, which means that the model is not valid.
2. The alternative hypothesis is the claim that **at least one** of the regression coefficients is not zero. The alternative hypothesis is the claim that at least one of the independent variables is linearly related to the dependent variable, which means that the model is valid. The alternative hypothesis does not say that all of the regression coefficients are not zero, only that at least one of them is not zero. The alternative hypothesis does not tell us which independent variables are related to the dependent variable.
3. The **p — value** for the **overall model test** is located in the middle part of the table under the **Significance F column** heading in the **Regression row** (right underneath the **ANOVA heading**). You will notice a **p — value** column heading at the bottom of the table in the rows corresponding to the independent variables. These **p — value** in the bottom part of the table are not related to the overall model test we are conducting here. These **p — value** in the independent variable rows are the **p — value** we will need when we conduct tests on the individual regression coefficients in the next section.
4. The **p — value** of **0.0017** is a small probability compared to the significance level, and so is unlikely to happen, assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, at least one of the regression coefficients is not zero, and at least one independent variable is linearly related to the dependent variable.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats2ed/?p=317#oembed-1>

Video: “Basic Excel Business Analytics #51: Testing Significance of Regression Relationship with p-value” by excelisfun [20:45] is licensed under the Standard YouTube License. *Transcript and closed captions available on YouTube.*

Exercises

1. A local restaurant advocacy group wants to study the relationship between a restaurant’s average weekly profit, the restaurant’s seating capacity, and the average daily traffic that passes the restaurant’s location. The group took a sample of restaurants and recorded their average weekly profit (in \$1000s), the seating restaurant’s seating capacity, and the average number of cars (in 1000s) that passes the restaurant’s location. The data is recorded in the following table:

Seating Capacity	Traffic Count (1000s)	Weekly Net Profit (\$1000s)
120	19	23.8
180	8	29.2
150	12	22
180	15	26.2
220	16	33.5
235	10	32
115	18	22.4
110	12	20.4
165	21	23.7
220	20	34.7
140	24	27.1
145	24	23.3
140	13	20.9
200	14	29.6
210	14	31.4
175	12	23.2
175	15	31.1
190	17	28.2
100	23	25.2
145	20	20.7
135	13	37.2
25	13	26.3
140	25	20
130	14	28.2
135	10	24.6
160	23	23.7

In Question 1 of Section 13.1, we found the regression model to predict the average weekly profit from other variables. At the 5% significance level, test the validity of the model.

Click to see Answer

- Hypotheses: $H_0 : \beta_1 = \beta_2$
 $H_a : \text{at least one } \beta_i \text{ is not } 0$
- $p - \text{value} = 0.0205$
- At the 5\% significance level, there is enough evidence to suggest that there is a relationship between the dependent variable “weekly profit” and the set of independent variables “seating capacity” and “traffic count.”

2. A local university wants to study the relationship between a student’s GPA, the average number of hours they spend studying each night, and the average number of nights they go out each week. The university took a sample of students and recorded the following data:

GPA	Average Number of Hours Spent Studying Each Night	Average Number of Nights Go Out Each Week
3.72	5	1
3.88	3	1
3.67	2	1
3.87	3	4
2.49	1	4
1.29	1	2
1.01	2	4
2.12	1	1
1.9	1	5
3.42	3	2
1.33	1	4
1.07	0	2
2.75	3	1
3.82	4	1
3.91	5	0
2.25	2	3
2.06	1	5
2.92	3	2
3.06	3	1
3.65	2	2
3.69	4	1

In Question 2 of Section 13.1, we found the regression model to predict GPA from other variables. At the 1\% significance level, test the validity of the model.

Click to see Answer

- Hypotheses: $H_0 : \beta_1 = \beta_2$
 $H_a : \text{at least one } \beta_i \text{ is not } 0$
- $p - \text{value} = 0.0002$
- At the 1\% significance level, there is enough evidence to suggest that there is a relationship between the dependent variable “GPA” and the set of independent variables

“average number of hours spent studying each night” and “average number of nights go out each week.”

3. A very large company wants to study the relationship between the salaries of employees in management positions, their age, the number of years the employee spent in college, and the number of years the employee has been with the company. A sample of management employees is taken, and the data is recorded below:

Age	Years of College	Years with Company	Salary (\$1000s)
60	8	29	317.3
33	3	5	97.3
57	6	27	263.1
32	4	5	101.3
31	6	3	114.2
61	8	19	350.4
41	7	8	146.9
35	4	2	91.7
51	6	21	198.2
50	8	10	196.5
57	5	15	105.7
49	6	18	118.3
62	7	27	305.2
52	8	26	239.9
39	4	8	145.9
42	7	5	175.4
62	4	24	219.4
60	4	22	202.1
65	3	21	196.3
40	4	10	143.9
62	6	29	408.7
53	7	5	145.2
48	8	5	175.1
61	5	6	152.7
38	7	3	99.7

40	7	12	174.9
45	7	7	149.2
58	7	14	282.8
38	4	3	95.7
41	5	18	232.8

In Question 3 of Section 13.1, we found the regression model to predict salary from other variables. At the 1\% significance level, test the validity of the model.

Click to see Answer

- Hypotheses: $H_0 : \beta_1 = \beta_2 = \beta_3$
 $H_a : \text{at least one } \beta_i \text{ is not } 0$
- $p - \text{value} = 0.0000002$
- At the 1\% significance level, there is enough evidence to suggest that there is a relationship between the dependent variable “salary” and the set of independent variables “age”, “years of college”, and “years at company.”

“13.5 Testing the Significance of the Overall Model” and “13.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

13.5 TESTING THE REGRESSION COEFFICIENTS

LEARNING OBJECTIVES

- Conduct and interpret a hypothesis test on individual regression coefficients.

Previously, we learned that the population model for the multiple regression equation is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

where x_1, x_2, \dots, x_k are the independent variables, $\beta_0, \beta_1, \dots, \beta_k$ are the population parameters of the regression coefficients, and ϵ is the error variable. In multiple regression, we estimate each population regression coefficient β_i with the sample regression coefficient b_i .

In the previous section, we learned how to conduct an overall model test to determine if the regression model is valid. If the outcome of the overall model test is that the model is valid, then at least one of the independent variables is related to the dependent variable—in other words, at least one of the regression coefficients β_i is not zero. However, the overall model test does not tell us which independent variables are related to the dependent variable. To determine which independent variables are related to the dependent variable, we must test each of the regression coefficients.

Testing the Regression Coefficients

For an individual regression coefficient, we want to test if there is a relationship between the dependent variable y and the independent variable x_i .

- **No Relationship.** There is no relationship between the dependent variable y and the independent variable x_i . In this case, the regression coefficient β_i is zero. This is the claim for the null hypothesis in an individual regression coefficient test: $H_0 : \beta_i = 0$.
- **Relationship.** There is a relationship between the dependent variable y and the independent variable x_i . In this case, the regression coefficient β_i is not zero. This is the claim for the alternative hypothesis in an individual regression coefficient test: $H_a : \beta_i \neq 0$. We are not interested if the regression coefficient β_i is positive or negative, only that it is not zero. We only need to find out if the regression coefficient is not zero to demonstrate that there is a relationship between the dependent variable and the independent variable. This makes the test on a regression coefficient a two-tailed test.

In order to conduct a hypothesis test on an individual regression coefficient β_i , we need to use the distribution of the sample regression coefficient b_i :

- The mean of the distribution of the sample regression coefficient is the population regression coefficient β_i .
- The standard deviation of the distribution of the sample regression coefficient is σ_{b_i} . Because we do not know the population standard deviation, we must estimate σ_{b_i} with the sample standard deviation s_{b_i} .
- The distribution of the sample regression coefficient follows a normal distribution.

Because we are using a sample standard deviation to estimate a population standard deviation in a normal distribution, we need to use a t -distribution with $n - k - 1$ degrees of freedom to find the p -value for the test on an individual regression coefficient. The t -score for the test is

$$t = \frac{b_i - \beta_i}{s_{b_i}}.$$

Conducting a Hypothesis Test on a Regression Coefficient

Follow these steps to perform a hypothesis test on an individual regression coefficient:

1. Write down the null hypothesis that there is no relationship between the dependent variable y and the independent variable x_i :

$$H_0 : \beta_i = 0$$

2. Write down the alternative hypotheses that there is a relationship between the dependent variable y and the independent variable x_i :

$$H_a : \beta_i \neq 0$$

3. Collect the sample information for the test and identify the significance level α .
4. The p – value is the sum of the area in the tails of the t -distribution. The t -score and degrees of freedom are

$$t = \frac{b_i - \beta_i}{s_{b_i}}$$

$$df = n - k - 1$$

5. Compare the p – value to the significance level and state the outcome of the test.
 - If p – value $\leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If p – value $> \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
6. Write down a concluding sentence specific to the context of the question.

The required t -score and p – value for the test can be found on the regression summary table, which we learned how to generate in Excel in a previous section.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken, and the data is recorded in the table below. The employee's income is recorded in \$1000s, and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60

7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

At the 5% significance level, test the relationship between the dependent variable “job satisfaction” and the independent variable “hours of unpaid work per week”.

Solution

Hypotheses:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

p – value:

The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The p – value for the test on the hours of unpaid work per week regression coefficient is in the bottom part of the table under the **P-value column** of the **Hours of Unpaid Work per Week** row. So the p – value = 0.0082.

Conclusion:

Because p – value = 0.0082 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level, there is enough evidence to suggest that there is a relationship between the dependent variable “job satisfaction” and the independent variable “hours of unpaid work per week.”

NOTES

1. The null hypothesis $\beta_1 = 0$ is the claim that the regression coefficient for the independent variable x_1 is zero. That is, the null hypothesis is the claim that there is no relationship between the dependent variable and the independent variable, “hours of unpaid work per week.”
2. The alternative hypothesis is the claim that the regression coefficient for the independent variable x_1 is not zero. The alternative hypothesis is the claim that there is a relationship between the dependent variable and the independent variable, “hours of unpaid work per week.”
3. When conducting a test on a regression coefficient, make sure to use the correct subscript on β to correspond to how the independent variables were defined in the regression model and which independent variable is being tested. Here the subscript on β is 1 because the “hours of unpaid work per week” is defined as x_1 in the regression model.
4. The **p — value** for the tests on the regression coefficients are located in the bottom part of the table under the **P-value column** heading in the corresponding independent variable row.
5. Because the alternative hypothesis is a \neq , the **p — value** is the sum of the area in the tails of the t -distribution. This is the value calculated out by Excel in the regression summary table.
6. The **p — value** of 0.0082 is a small probability compared to the significance level and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the regression coefficient β_1 is not zero, and so there is a relationship between the dependent variable “job satisfaction” and the independent variable “hours of unpaid work per week.” This means that the independent variable “hours of unpaid work per week” is useful in predicting the dependent variable.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken, and the data is recorded in the table below. The employee's income is recorded in \$1000s, and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60

7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

At the 5% significance level, test the relationship between the dependent variable “job satisfaction” and the independent variable “age”.

Solution

Hypotheses:

$$H_0 : \beta_2 = 0$$

$$H_a : \beta_2 \neq 0$$

p – value:

The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The p – value for the test on the age regression coefficient is in the bottom part of the table under the **P-value column** of the **Age row**. So the p – value = 0.8439.

Conclusion:

Because p – value = 0.8439 > 0.05 = α , we do not reject the null hypothesis. At the 5\% significance level, there is not enough evidence to suggest that there is a relationship between the dependent variable “job satisfaction” and the independent variable “age.”

NOTES

1. The null hypothesis $\beta_2 = 0$ is the claim that the regression coefficient for the independent variable x_2 is zero. That is, the null hypothesis is the claim that there is no relationship between the dependent variable and the independent variable “age.”
2. The alternative hypothesis is the claim that the regression coefficient for the independent variable x_2 is not zero. The alternative hypothesis is the claim that there is a relationship between the dependent variable and the independent variable “age.”
3. When conducting a test on a regression coefficient, make sure to use the correct subscript on β to correspond to how the independent variables were defined in the regression model and which independent variable is being tested. Here, the subscript on β is 2 because “age” is defined as x_2 in the regression model.
4. The p — value of 0.8439 is a large probability compared to the significance level and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the regression coefficient β_2 is zero, and so there is no relationship between the dependent variable “job satisfaction” and the independent variable “age.” This means that the independent variable “age” is not particularly useful in predicting the dependent variable.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee’s job satisfaction from the number of hours of unpaid work per week the employee does, the employee’s age, and the employee’s income. A sample of 25 employees at the company is taken, and the data is recorded in the table below. The employee’s income is recorded in \$1000s, and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60

7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

At the 5% significance level, test the relationship between the dependent variable “job satisfaction” and the independent variable “income”.

Solution

Hypotheses:

$$H_0 : \beta_3 = 0$$

$$H_a : \beta_3 \neq 0$$

p – value:

The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	>df	SS	MS	F	Significance F	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The p – value for the test on the income regression coefficient is in the bottom part of the table under the **P-value column** of the **Income row**. So the p – value = 0.0060.

Conclusion:

Because p – value = 0.0060 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5\% significance level, there is enough evidence to suggest that there is a relationship between the dependent variable “job satisfaction” and the independent variable “income.”

NOTES

1. The null hypothesis $\beta_3 = 0$ is the claim that the regression coefficient for the independent variable x_3 is zero. That is, the null hypothesis is the claim that there is no relationship between the dependent variable and the independent variable “income.”
2. The alternative hypothesis is the claim that the regression coefficient for the independent variable x_3 is not zero. The alternative hypothesis is the claim that there is a relationship between the dependent variable and the independent variable “income.”
3. When conducting a test on a regression coefficient, make sure to use the correct subscript on β to correspond to how the independent variables were defined in the regression model and which independent variable is being tested. Here, the subscript on β is 3 because “income” is defined as x_3 in the regression model.
4. The p — value of 0.0060 is a small probability compared to the significance level and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the regression coefficient β_3 is not zero, and so there is a relationship between the dependent variable “job satisfaction” and the independent variable “income.” This means that the independent variable “income” is useful in predicting the dependent variable.

Exercises

1. A local restaurant advocacy group wants to study the relationship between a restaurant’s average weekly profit, the restaurant’s seating capacity, and the average daily traffic that passes the restaurant’s location. The group took a sample of restaurants and recorded their average weekly profit (in \$1000s), the seating restaurant’s seating capacity, and the average number of cars (in 1000s) that passes the restaurant’s location. The data is recorded in the following table:

Seating Capacity	Traffic Count (1000s)	Weekly Net Profit (\$1000s)
120	19	23.8
180	8	29.2
150	12	22
180	15	26.2
220	16	33.5
235	10	32
115	18	22.4
110	12	20.4
165	21	23.7
220	20	34.7
140	24	27.1
145	24	23.3
140	13	20.9
200	14	29.6
210	14	31.4
175	12	23.2
175	15	31.1
190	17	28.2
100	23	25.2
145	20	20.7
135	13	37.2
25	13	26.3
140	25	20
130	14	28.2
135	10	24.6
160	23	23.7

In Question 1 of Section 13.1, we found the regression model to predict the average weekly profit from other variables.

- a. At the 5\% significance level, test the coefficient of seating capacity.
- b. At the 5\% significance level, test the coefficient of traffic count.

Click to see Answer

- Hypotheses: $H_0 : \beta_1 = 0$
 $H_a : \beta_1 \neq 0$
- $p - \text{value} = 0.0144$
- At the 5\% significance level, there is enough evidence to suggest that there is a relationship between the dependent variable “weekly profit” and the independent variable “seating capacity”.

- Hypotheses: $H_0 : \beta_2 = 0$
 $H_a : \beta_2 \neq 0$
- $p - \text{value} = 0.2645$
- At the 5\% significance level, there is not enough evidence to suggest that there is a relationship between the dependent variable “weekly profit” and the independent variable “traffic count”.

2. A local university wants to study the relationship between a student’s GPA, the average number of hours they spend studying each night, and the average number of nights they go out each week. The university took a sample of students and recorded the following data:

GPA	Average Number of Hours Spent Studying Each Night	Average Number of Nights Go Out Each Week
3.72	5	1
3.88	3	1
3.67	2	1
3.87	3	4
2.49	1	4
1.29	1	2
1.01	2	4
2.12	1	1
1.9	1	5
3.42	3	2
1.33	1	4
1.07	0	2
2.75	3	1
3.82	4	1
3.91	5	0
2.25	2	3
2.06	1	5
2.92	3	2
3.06	3	1
3.65	2	2
3.69	4	1

In Question 2 of Section 13.1, we found the regression model to predict GPA from other variables.

- At the 5\% significance level, test the coefficient of average number of hours spent studying each night.
- At the 5\% significance level, test the coefficient of the average number of nights go out each week.

Click to see Answer

- Hypotheses: $H_0 : \beta_1 = 0$
 $H_a : \beta_1 \neq 0$
- $p - \text{value} = 0.0009$
- At the 1\% significance level, there is enough evidence to suggest that there is a relationship between the dependent variable “GPA” and the independent variable “average number of hours spent studying each night”.

- Hypotheses: $H_0 : \beta_2 = 0$
 $H_a : \beta_2 \neq 0$
- $p - \text{value} = 0.5083$
- At the 1\% significance level, there is not enough evidence to suggest that there is a relationship between the dependent variable “GPA” and the independent variable “average number of nights go out each week”.

3. A very large company wants to study the relationship between the salaries of employees in management positions, their age, the number of years the employee spent in college, and the number of years the employee has been with the company. A sample of management employees is taken, and the data is recorded below:

Age	Years of College	Years with Company	Salary (\$1000s)
60	8	29	317.3
33	3	5	97.3
57	6	27	263.1
32	4	5	101.3
31	6	3	114.2
61	8	19	350.4
41	7	8	146.9
35	4	2	91.7
51	6	21	198.2
50	8	10	196.5
57	5	15	105.7
49	6	18	118.3
62	7	27	305.2
52	8	26	239.9
39	4	8	145.9
42	7	5	175.4
62	4	24	219.4
60	4	22	202.1
65	3	21	196.3
40	4	10	143.9
62	6	29	408.7
53	7	5	145.2
48	8	5	175.1
61	5	6	152.7
38	7	3	99.7

40	7	12	174.9
45	7	7	149.2
58	7	14	282.8
38	4	3	95.7
41	5	18	232.8

In Question 3 of Section 13.1, we found the regression model to predict salary from other variables.

- At the 1\% significance level, test the coefficient of age.
- At the 1\% significance level, test the coefficient of years of college.
- At the 1\% significance level, test the coefficient of years with the company.

Click to see Answer

- Hypotheses: $H_0 : \beta_1 = 0$
 $H_a : \beta_1 \neq 0$
 - $p - \text{value} = 0.2383$
 - At the 1\% significance level, there is not enough evidence to suggest that there is a relationship between the dependent variable “salary” and the independent variable “age”.

 - Hypotheses: $H_0 : \beta_2 = 0$
 $H_a : \beta_2 \neq 0$
 - $p - \text{value} = 0.097$
 - At the 1\% significance level, there is enough evidence to suggest that there is a relationship between the dependent variable “salary” and the independent variable “years of college”.

 - Hypotheses: $H_0 : \beta_3 = 0$
 $H_a : \beta_3 \neq 0$
 - $p - \text{value} = 0.0005$
 - At the 1\% significance level, there is enough evidence to suggest that there is a relationship between the dependent variable “salary” and the independent variable “years with company”.
-

“13.6 Testing the Regression Coefficients” and “13.8 Exercises” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

13.6 MULTICOLLINEARITY

LEARNING OBJECTIVES

- Define multicollinearity and explain its impact on multiple regression.

The term independent variable applies to any variable that is used to predict or explain the value of the dependent variable. But this does not mean that the independent variables themselves are unrelated to each other. In fact, most independent variables in multiple regression models share some degree of relatedness. For example, if “distance travelled” and “litres of gas consumed” are the independent variables in a regression model to predict the dependent variable “travel time,” the variables “distance travelled” and “litres of gas consumed” are highly correlated.

When two or more independent variables in a regression model are highly correlated to each other, **multicollinearity** exists between the independent variables. Consequently, the conclusions about the relationship between the dependent variable and the individual independent variables may be affected when the independent variables are related to each other. In addition, multicollinearity may affect the outcome of the tests on the individual regression coefficients. But multicollinearity does not affect the outcome of the overall test on the regression model.

Even though the overall model test may conclude that there is a relationship between the dependent variable and the set of independent variables, multicollinearity amongst the independent variables may cause all of the tests on the individual regression coefficients to conclude that none of the individual independent variables are related to the dependent variable. One way to address the problem of multicollinearity is to avoid including independent variables that are highly correlated or remove one of two highly correlated independent variables from the model.

“13.7 Multicollinearity” from Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

PART XIV

TIME SERIES ANALYSIS

In many situations, we need to explore and model how data changes over time, which allows us to make predictions about what will happen in the future. In particular, businesses are interested in forecasting future needs and demands. For example, a company that produces a certain product needs accurate predictions about their volume of sales in the near future in order to make decisions about things such as production schedules, the purchasing of raw materials, inventory policies, and sales quotas. A good forecast ensures the company is operating efficiently and cost-effectively. A poor forecast may result in poor planning and increased costs.

In time series analysis, data is collected over a period of time at regular intervals, such as monthly, quarterly, or yearly. Using this time-based data, a model can be built to predict what will happen in future time intervals. For example, a business might collect quarterly sales revenue over a three-year period and then use that data to build a model to predict the sales revenue for each quarter of the fourth year. There are several different time series models, but some models will give more accurate predictions than others depending on the traits in the particular time series data. When using time series models, we need to understand how the different models make their forecasts and identify which model to use on a given time series to ensure the most accurate predictions from the model.

CHAPTER OUTLINE

14.1 Time Series Patterns

14.2 Measures of Forecast Accuracy

14.3 Smoothing Models

14.4 Seasonal Indices

14.5 Regression Models

14.1 TIME SERIES PATTERNS

LEARNING OBJECTIVES

- Identify and describe the components of a time series data.
- Construct a time series graph.
- Analyze and interpret time series data presented in a time series graph.

Suppose that we want to study the temperature range of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with this data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. However, all of these methods ignore a portion of the data that we have collected—time. Because each date is paired with the temperature reading for that day, we do not have to think of the data as being random. Instead, we can use the times given to impose a chronological order on the data.

A **time series** is a set of observations for a particular variable taken at regular intervals of time or over successive periods of time. The interval in a time series may be taken every second, every minute, every hour, every day, every week, every month, every year, or at any other regular time interval. The data in a time series must be indexed in chronological order in order to spot patterns or trends that occur in the data over time. If these trends or patterns continue into the future, we can use the past pattern to create a model and make predictions about what will happen in the future.

EXAMPLE

The data below records the daily high temperature, in Celsius, every day over a two-week period.

Day	Temperature (Celsius)
1	6
2	5
3	3
4	5
5	2
6	6
7	9
8	10
9	4
10	5
11	10
12	3
13	4
14	8

This is a time series where the data points (temperature) are measured daily over a period of time. The interval of time for the time series is daily.

Time Series Graphs

A **time series graph**, also called a **time series plot**, is a useful tool to help us identify any underlying patterns in the time series data. A time series graph is a graphical presentation of the relationship between time (in chronological order) and the time series variable. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of

the variable that we are measuring. By doing this, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

CONSTRUCTING A TIME SERIES GRAPH IN EXCEL

To create a time series graph in Excel:

1. Time is the horizontal axis, and the quantity being measured is the vertical axis.
2. If necessary, rearrange the columns so that the column containing time is on the left and the measured variable is on the right. (Excel always places the variable on the left on the horizontal axis.)
3. Go to the **Insert** tab. In the **Charts** group, click on **Scatter**. Select the scatter diagram with the markers (points) connected by **straight** lines.
4. Using the chart tools, add axis titles, including both the variable names and units on the axes.
5. Using the chart tools, add a chart title.

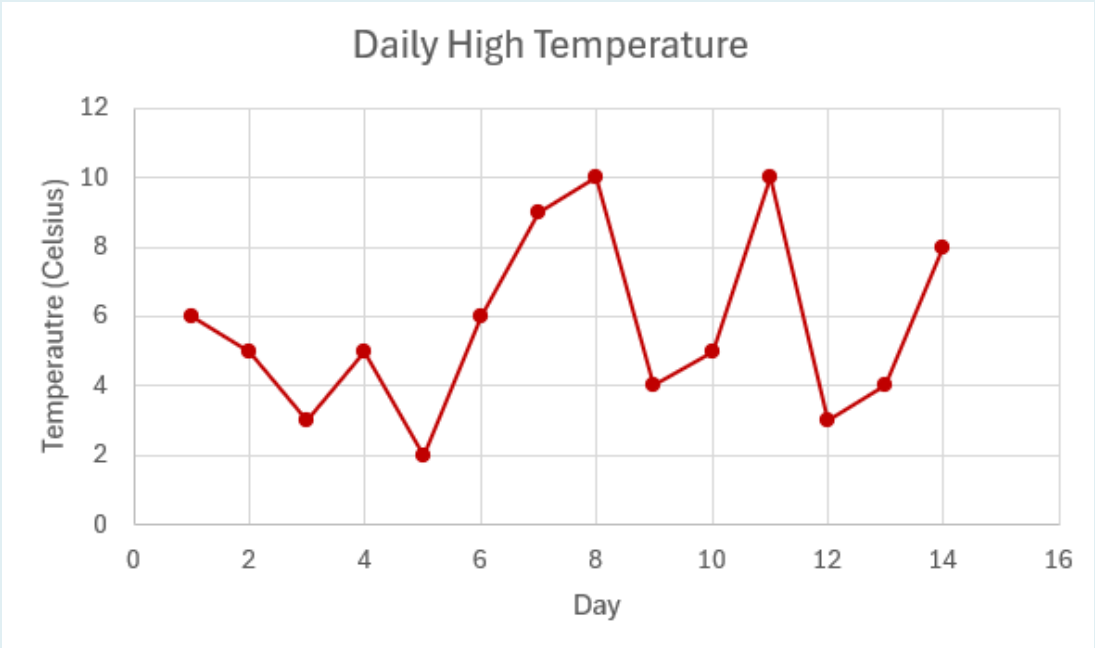
Visit the Microsoft page for more information about creating a time series graph in Excel.

EXAMPLE

The time series below records the daily high temperature, in Celsius, every day over a two-week period. Construct a time series graph for this data.

Day	Temperature (Celsius)
1	6
2	5
3	3
4	5
5	2
6	6
7	9
8	10
9	4
10	5
11	10
12	3
13	4
14	8

Solution

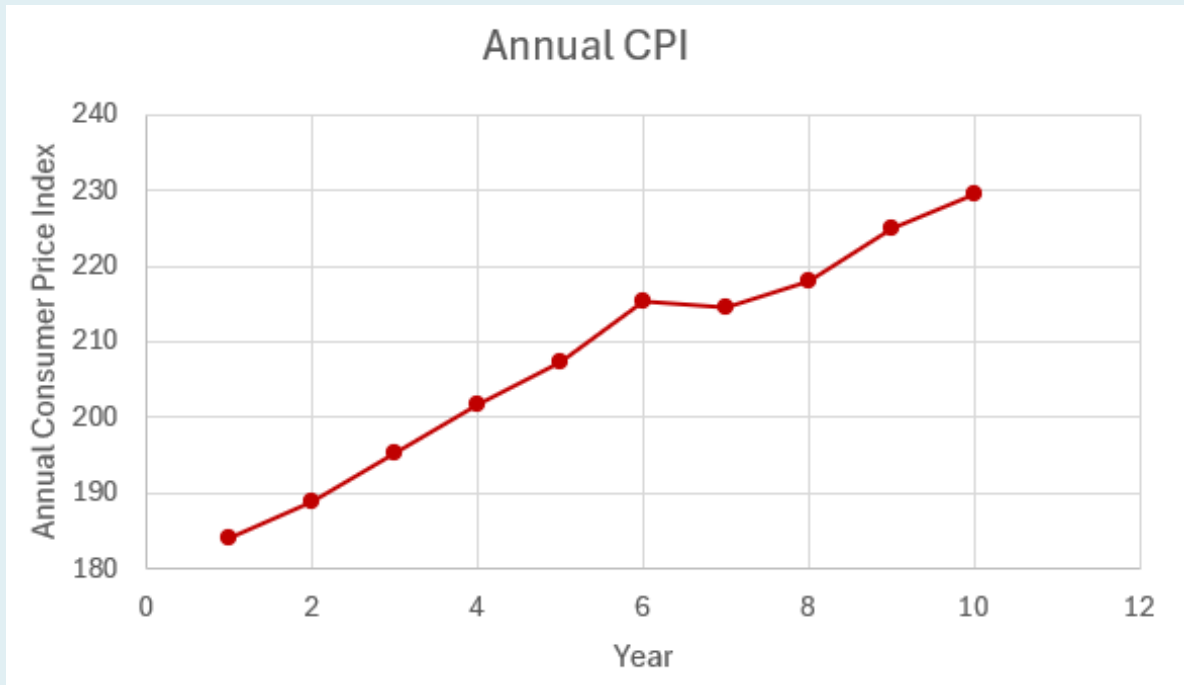


EXAMPLE

The following data shows the Annual Consumer Price Index each year for ten years. Construct a time series graph for this data.

Year	Annual Consumer Price Index
1	184.0
2	188.9
3	195.3
4	201.6
5	207.342
6	215.303
7	214.537
8	218.056
9	224.939
10	229.594

Solution



Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

Components of a Time Series

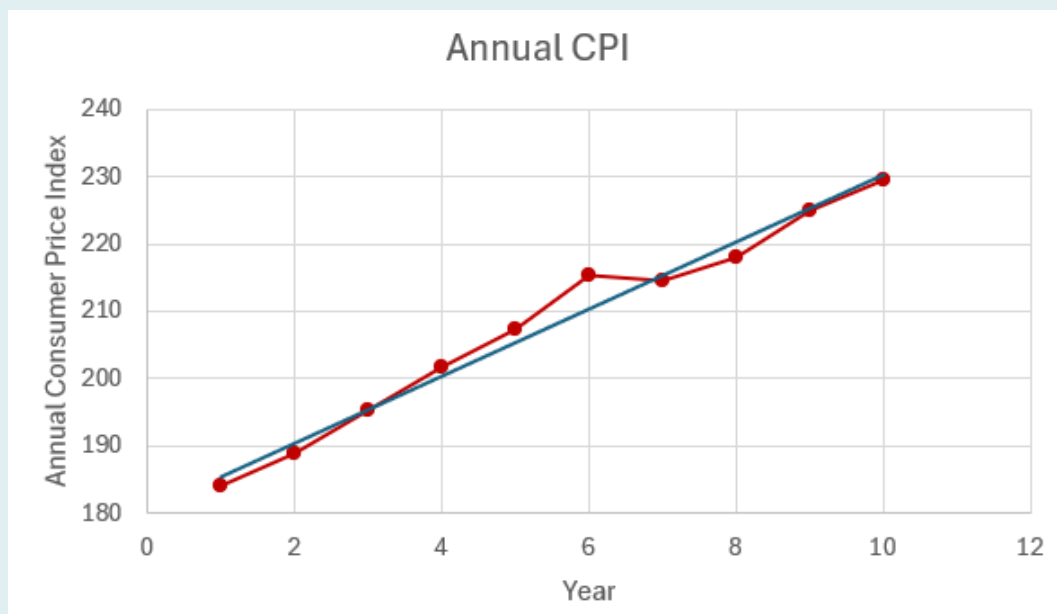
Time series forecasting models use historical time series data to develop a forecast of the future. Although most models can be applied to any time series data, the choice of model for a particular time series depends on the patterns or components present in the time series. Some models are better at capturing certain components than other models. To get the most accurate predictions, we want to select a model that is good at modelling the components present in the given time series. By analyzing the graph of a time series, we can often identify the components present in the time series, allowing us to break down a complex time series into more manageable parts in order to build better models and forecasts. Time series data typically exhibits one or more of four possible components that help us understand the underlying patterns present in a time series.

Trend Pattern

A **trend pattern** is the long-term movement or general direction of the data over a period of time. A trend could be upward, downward, or a flat trend and is generally the result of long-term factors or influences on the data. For example, in a time series for the price of a stock the overall increase or decrease in value of the stock over several years is considered a trend. Often, we look for linear trends in the data, and when a linear trend is present, we can use a trendline or simple linear regression model to model the data. But trends do not have to follow a linear model. Other examples of trends in a time series include quadratic trends or exponential trends.

EXAMPLE

The following time series graph shows the Annual Consumer Price Index each year for ten years. The trendline, found through simple linear regression, is included on the graph.



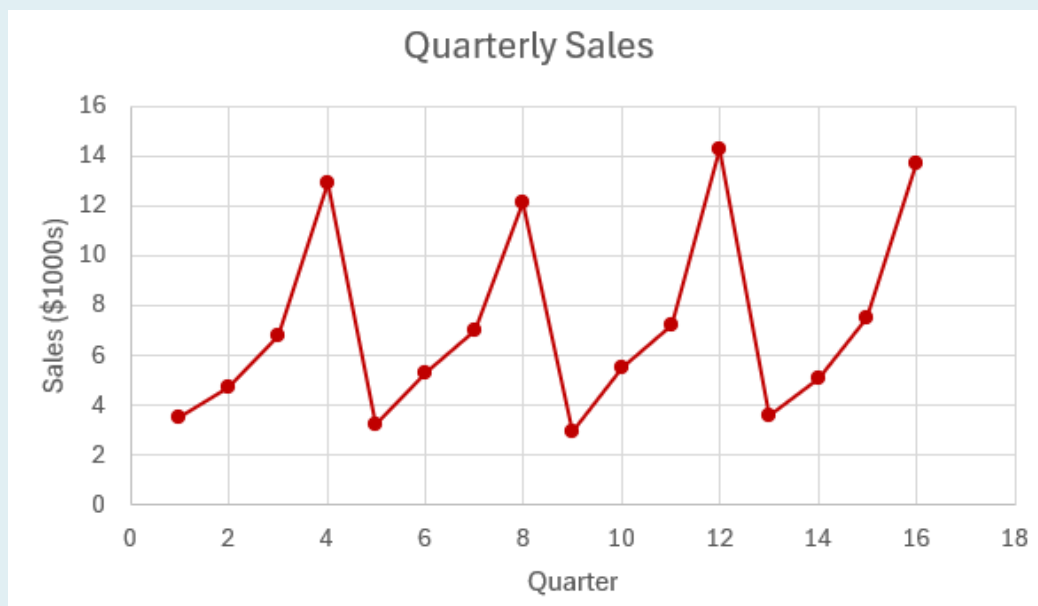
From the time series graph we can see the overall trend of the annual consumer price index is an increase over time. In this instance, the trend appears to be linear, which we can see by how closely the data follows the trendline.

Seasonality or Seasonal Patterns

Seasonal patterns are repeated patterns that occur over a one-year period due to seasonal influences. Repeated and predictable highs and lows in the time series data that occur at the same time each year indicate seasonality in the data. For example, retail sales experience predictable highs during the holiday season every year, which creates a seasonal pattern.

EXAMPLE

The following time series graph shows a business's sales each quarter for four years.



From the time series, we see seasonal patterns in the quarterly sales. We can see regular and predictable highs in every fourth quarter (quarters 4, 8, 12, and 16), which correspond to the increased sales the business sees during the holiday season in the last quarter of each year. We can also see regular and predictable lows in every first quarter (quarters 1, 5, 9, and 13), which correspond to the decrease in sales the business sees after the holiday season in the first quarter of the year.

Cycles or Cyclic Patterns

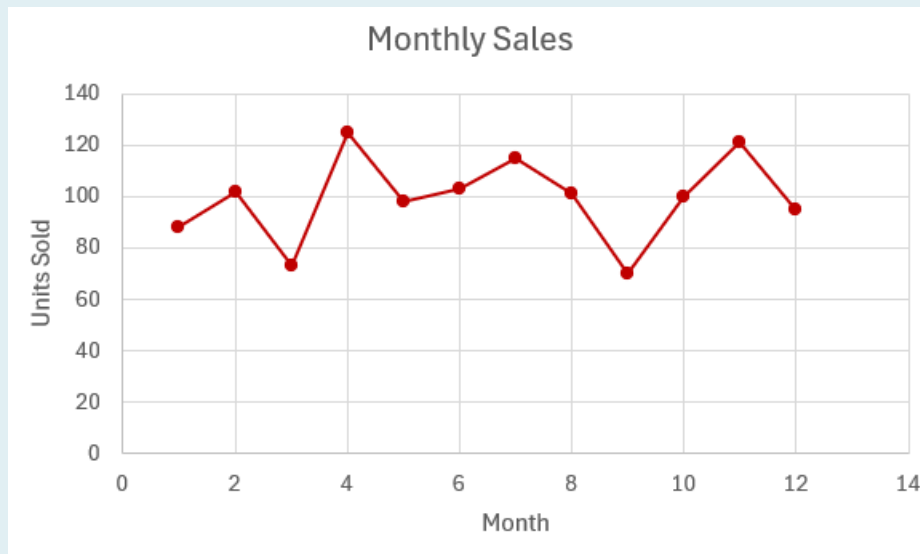
Unlike seasonality, which occurs at regular intervals, **cyclic patterns** occur at irregular intervals and are often influenced by multi-year economic or business cycles. These fluctuations are much longer in duration, typically spanning multiple years. Because of the extended time frame required for cycles, identifying a cyclic pattern in a time series is challenging.

Irregular Variation or Random Fluctuations

Random variation or **irregular components** in a time series are unpredictable or unforeseen changes in the data that cannot be attributed to any of the other components. Random fluctuations are often unexplained, reflecting the inherent randomness and unpredictability in the system.

EXAMPLE

The following time series graph shows the number of units sold for a particular product each month for one year.



There is no general trend or pattern to the data from left to right across the graph, so the time series

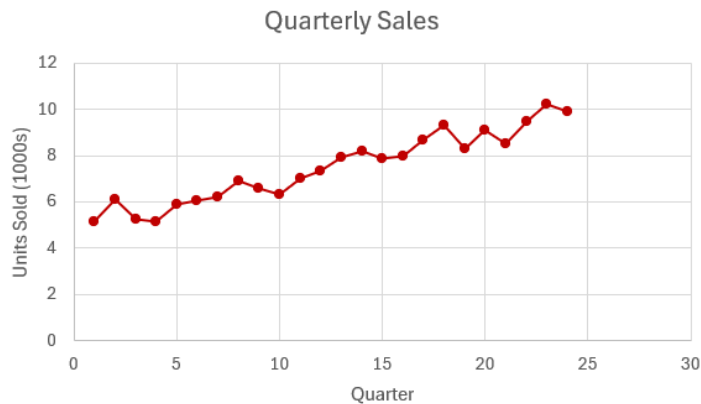
does not appear to have a trend component. With only one year of data, it is not possible to spot any seasonal effects or cyclical patterns. The time series graph does exhibit random variations.

Exercises

- For each of the following time series, identify the components of the time series present in the data.



a.



b.



c.

Click to see Answer

- a. seasonal effects, random variation
- b. trend, random variation
- c. seasonal effects, trend, random variation

2. The number of pizzas sold each week at a local pizza restaurant is given in the table below.

Week	Number of Pizzas Sold
1	120
2	135
3	140
4	115
5	130
6	145
7	160
8	170
9	125
10	130
11	150
12	140
13	155
14	180
15	165
16	170
17	190

- Create a time series graph for this data.
- What components of a time series are present in the time series?

Click to see Answer



-
- trend, random variation

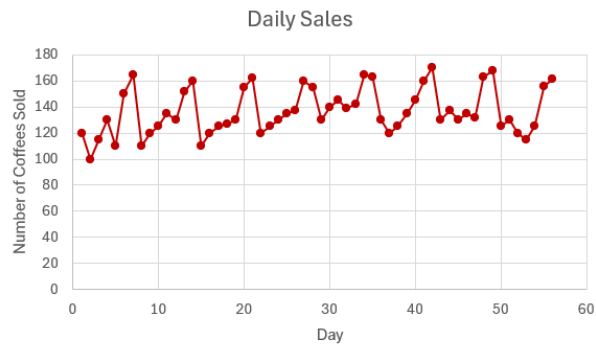
3. The number of coffees sold each day at a local coffee shop is given in the table below.

Day	Number of Coffees Sold
1	120
2	100
3	115
4	130
5	110
6	150
7	165
8	110
9	120
10	125
11	135
12	130
13	152
14	160
15	110
16	120
17	125
18	127
19	130
20	155
21	162
22	120
23	125
24	130
25	135
26	137
27	160
28	155
29	130

30	140
31	145
32	139
33	142
34	165
35	163
36	130
37	120
38	125
39	135
40	145
41	160
42	170
43	130
44	137
45	130
46	135
47	132
48	163
49	168
50	125
51	130
52	120
53	115
54	125
55	156
56	161

- Create a time series graph for this data.
- What components of a time series are present in the time series?

Click to see Answer



- a.
- b. seasonal effects, random variation

4. The quarterly demand, in 1000s, for a particular product is recorded in the table below.

Quarter	Demand (in 1000s)
1	50
2	75
3	85
4	70
5	60
6	95
7	100
8	85
9	70
10	115
11	125
12	110
13	80
14	140
15	145
16	130
17	90
18	165
19	175
20	155

- Create a time series graph for this data.
- What components of a time series are present in the time series?

Click to see Answer



a.

- b. trend, seasonal effects, random variation

Attribution

“2.2 Histograms, Frequency Polygons, and Time Series Graphs“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

14.2 MEASURES OF FORECAST ACCURACY

LEARNING OBJECTIVES

- Calculate and interpret forecast errors, including mean absolute deviation, mean square error, and mean absolute percent error.

Using time series data, we want to build a model from that data in order to make forecasts and predictions about the future. There are many different time series models, each giving a different forecast for the same data. But how accurate is the forecast obtained from a time series model, and how can we tell which of the models will give us the most accurate forecast?

Forecast accuracy is measured through the errors in the forecast. We will look at three different error analysis techniques—mean absolute deviation (MAD), mean square error (MSE), and mean absolute percent error ($MAPE$). Each of these techniques examines the errors in the forecast in slightly different ways, but they all involve working with the **forecast error**, which is the difference between the actual value in the data and the value forecasted by the model.

$$\text{Forecast Error} = \text{Actual Value} - \text{Forecast Value}$$

Mean Absolute Deviation

The **mean absolute deviation**, denoted MAD , is the average of the absolute value of the forecasting errors.

$$MAD = \frac{\sum |\text{forecast error}|}{\text{number of forecast errors}}$$

To calculate the mean absolute deviation:

1. Calculate the forecast errors.
2. Calculate the absolute value of each forecast error.
3. Sum up the absolute values of the errors found in Step 2.
4. Divide the sum in Step 3 by the number of errors.

The mean absolute deviation tells us the average difference between the actual values and the forecast values. In general, the smaller the mean absolute deviation, the better the model is at forecasting. A small MAD indicates that the average error is small, which means that, on average, the actual values and the forecast values are close in value.

NOTES

1. There will be an error for each time period that has **both** an actual value **and** a forecast value. If a time period has only an actual value and no forecast value, then there is no error for that time period. Similarly, if a time period has only a forecast value and no actual value, then there is no error for that time period.
2. The mean absolute deviation for a time series is similar to the standard deviation for a set of data. The standard deviation for a set of data tells us the average deviation of the data from the mean. The mean absolute deviation tells us the average deviation of the time series from the forecast.
3. Why are we calculating the average of the **absolute value** of the errors and not the average of the errors themselves? Some of the forecast errors will be positive (when the actual value is greater than the forecast value), and some forecast errors will be negative (when the actual value is less than the forecast value). Consequently, the positive and negative errors have a tendency to cancel each other out, which means the average of the errors will be small regardless of how good or bad the model is at forecasting the time series. By taking the absolute value of the errors, the differences between the positive and negative errors are eliminated because the absolute values of the errors are all positive.

EXAMPLE

A local company produces and sells a certain product. The number of units sold each month for a year is recorded in the table below. The company created a forecast to predict the number of units sold each month.

Month	Actual Number of Units Sold	Forecasted Number of Units Sold
January	83	82
February	102	105
March	75	82
April	125	117
May	100	95
June	103	111
July	118	115
August	101	97
September	70	78
October	99	90
November	120	125
December	97	103

1. Calculate the mean absolute deviation for the forecast.
2. Interpret the mean absolute deviation.

Solution

1.

Month	Actual Number of Units Sold	Forecasted Number of Units Sold	Forecast Error	Forecast Error
January	83	82	1	1
February	102	105	−3	3
March	75	82	−7	7
April	125	117	8	8
May	100	95	5	5
June	103	111	−8	8
July	118	115	3	3
August	101	97	4	4
September	70	78	−8	8
October	99	90	9	9
November	120	125	−5	5
December	97	103	−6	6
			Sum	67

$$\begin{aligned}
 MAD &= \frac{\sum |\text{forecast error}|}{\text{number of forecast errors}} \\
 &= \frac{67}{12} \\
 &= 5.583
 \end{aligned}$$

2. On average, the forecasted values differ by 5.583 units sold from the actual values.

NOTES

1. To calculate the *MAD*, add up the absolute value of the errors column and divide the sum by the number of errors. In this example, there are 12 errors, so we divide the column sum by 12.
2. The mean absolute deviation is just the average (or mean) of the absolute values of the errors. After calculating the absolute value of the errors, calculate the mean of the absolute value of

the errors. In the above example, just calculate the mean of the absolute value of the errors column to get the *MAD*.

3. The units of the *MAD* are the same units as the time series variable. In this case, the time series variable is measuring the number of units sold, so the units of the *MAD* are also the number of units sold.
4. We interpret the *MAD* similar to how we interpret the standard deviation. The *MAD* tells us the average difference between the forecasted values and the actual values. When interpreting the *MAD*, be specific to the context of the question and include units with the *MAD*.

Mean Square Error

The **mean square error**, denoted *MSE*, is the average of the squared forecasting errors.

$$MSE = \frac{\sum(\text{forecast error})^2}{\text{number of forecast errors}}$$

To calculate the mean square error:

1. Calculate the forecast errors.
2. Calculate the square of each forecast error.
3. Sum up the squared errors found in Step 2.
4. Divide the sum in Step 3 by the number of errors.

In general, the smaller the mean square error, the better the model is at forecasting. A small *MSE* indicates that the average squared error is small, which means that, on average, the actual values and the forecast values are close in value.

NOTES

1. There will be an error for each time period that has **both** an actual value **and** a forecast value. If a time period has only an actual value and no forecast value, then there is no error for that time period. Similarly, if a time period has only a forecast value and no actual value, then there is no error for that time period.
2. The mean square error for a time series is similar to the variance for a set of data.
3. Why are we calculating the average of the **squared** errors? As noted in the discussion above about mean absolute deviation, some of the forecast errors will be positive and some forecast errors will be negative. Consequently, the positive and negative errors have a tendency to cancel each other out. By squaring the errors, the differences between the positive and negative errors are eliminated because all of the squared errors are positive.

EXAMPLE

A local company produces and sells a certain product. The number of units sold each month for a year is recorded in the table below. The company created a forecast to predict the number of units sold each month. Calculate the mean square error for the forecast.

Month	Actual Number of Units Sold	Forecasted Number of Units Sold
January	83	82
February	102	105
March	75	82
April	125	117
May	100	95
June	103	111
July	118	115
August	101	97
September	70	78
October	99	90
November	120	125
December	97	103

Solution

Month	Actual Number of Units Sold	Forecasted Number of Units Sold	Forecast Error	(Forecast Error) ²
January	83	82	1	1
February	102	105	−3	9
March	75	82	−7	49
April	125	117	8	64
May	100	95	5	25
June	103	111	−8	64
July	118	115	3	9
August	101	97	4	16
September	70	78	−8	64
October	99	90	9	81
November	120	125	−5	25
December	97	103	−6	36
			Sum	443

$$\begin{aligned}
 MSE &= \frac{\sum(\text{forecast error})^2}{\text{number of forecast errors}} \\
 &= \frac{443}{12} \\
 &= 36.917
 \end{aligned}$$

NOTES

1. To calculate the *MSE*, add up the squared errors column and divide the sum by the number of errors. In this example, there are **12** errors, so we divide the column sum by **12**.
2. The mean square error is just the average (or mean) of the squared errors. After calculating the squared errors, calculate the mean of the squared errors. In the above example, just calculate the mean of the squared errors column to get the *MSE*.
3. The units of the *MSE* are the squared units of the time series variable. In this case, the time series variable is measuring the number of units sold, so the units of the *MSE* are the

number of units sold squared.

4. Because the units of the MSE are squared units, it can be difficult to intuitively interpret the meaning of the MSE .

Mean Absolute Percent Error

Both the mean absolute deviation and the mean square error depend on the scale of the data, which makes it difficult to make comparisons between time series measured on different time intervals or between different time series. To make such comparisons, we need to compare the percent errors, which positions the errors on the same relative scale. The percent error is the forecast error divided by the actual value corresponding to that error.

$$\text{Percent Error} = \frac{\text{Forecast Error}}{\text{Actual Value}}$$

The **mean absolute percent error**, denoted $MAPE$, is the average of the absolute value of the percent errors.

$$\displaystyle{MAPE = \frac{\sum |\text{Percent Error}|}{\text{number of forecast errors}} \times 100\%}$$

To calculate the mean absolute percent error:

1. Calculate the forecast errors.
2. Divide each forecast error by its corresponding actual value.
3. Calculate the absolute value of the percent errors found in Step 2.
4. Sum up the absolute values of the percent errors found in Step 3.
5. Divide the sum in Step 4 by the number of errors.
6. Multiply by 100% to convert the result in step 5 to a percent.

The mean absolute percent error tells us the average percent difference between the actual values and the forecast values. In general, the smaller the mean absolute percent error, the better the

model is at forecasting. A small $MAPE$ indicates that the average percent error is small, which means that, on average, the actual values and the forecast values are close in value.

NOTES

1. There will be an error for each time period that has **both** an actual value **and** a forecast value. If a time period has only an actual value and no forecast value, then there is no error for that time period. Similarly, if a time period has only a forecast value and no actual value, then there is no error for that time period.
2. The mean absolute percent error tells us the average percent deviation of the time series values from the forecast values.
3. Why are we calculating the average of the **absolute value** of the percent errors and not the average of the percent errors themselves? As noted in the discussion above about mean absolute deviation, some of the forecast errors will be positive and some forecast errors will be negative. Consequently, the positive and negative errors have a tendency to cancel each other out. By taking the absolute value of the percent errors, the differences between the positive and negative errors are eliminated because the absolute value ensures all the percent errors are positive.

EXAMPLE

A local company produces and sells a certain product. The number of units sold each month for a year is recorded in the table below. The company created a forecast to predict the number of units sold each month.

Month	Actual Number of Units Sold	Forecasted Number of Units Sold
January	83	82
February	102	105
March	75	82
April	125	117
May	100	95
June	103	111
July	118	115
August	101	97
September	70	78
October	99	90
November	120	125
December	97	103

1. Calculate the mean absolute percent error for the forecast.
2. Interpret the mean absolute percent error.

Solution

1.

Month	Actual Number of Units Sold	Forecasted Number of Units Sold	Forecast Error	Percent Error
January	83	82	1	0.0120 ...
February	102	105	−3	0.0294 ...
March	75	82	−7	0.0933 ...
April	125	117	8	0.064
May	100	95	5	0.05
June	103	111	−8	0.0776 ...
July	118	115	3	0.0254 ...
August	101	97	4	0.0396 ...
September	70	78	−8	0.1142 ...
October	99	90	9	0.0909 ...
November	120	125	−5	0.0416 ...
December	97	103	−6	0.0618 ...
			Sum	0.7002 ...

$$\begin{aligned} \text{MAPE} &= \frac{\sum |\text{percent error}|}{\text{number of forecast errors}} \times 100\% \\ &= \frac{0.7002}{12} \times 100\% \\ &= 5.84\% \end{aligned}$$

2. On average, the forecasted values differ by 5.84% from the actual values.

NOTES

- To calculate the percent error, divide the forecast error by the actual value. In the above example, the percent error for January is $\frac{\text{forecast error}}{\text{actual value}} = \frac{1}{83} = 0.0120 \dots$
- To calculate the *MAPE*, add up the absolute value of the percent errors column, divide the sum by the number of errors and then multiply by 100 to convert the result to a percent. In this example, there are 12 errors, so we divide the column sum by 12.
- When calculating the *MAPE*, keep all of the decimals throughout the calculation to

prevent any round-off error.

4. The mean absolute percent error is just the average (or mean) of the absolute values of the percent errors. After calculating the absolute value of the percent errors, calculate the mean of the absolute value of the percent errors. In the above example, just calculate the mean of the absolute value of the percent errors column to get the *MAPE*.
5. The *MAPE* tells us the average percent difference between the forecasted values and the actual values.

Assessing Forecasts Using the Measures of Forecast Accuracy

We can use the measures of forecast accuracy to assess how good or bad the forecast is at modelling the known time series data. For all three measures of forecast accuracy, the smaller the value of the measure of forecast accuracy, the better the forecast is at modelling the data, and conversely, the larger the value of the measure of forecast accuracy, the worse the forecast is at modelling the data. For example, if a forecast produced a large *MAD*, then the forecast is probably a poor fit for the data.

Each of the above measures of forecast accuracy measure, albeit in different ways, how well the forecast is able to forecast the known values in the time series. But, we want to use the forecast to predict the values of a future time period where the actual value is unknown. How can we tell if this estimated value is accurate? In general, if a forecast works well on the known time series values and we expect the pattern present in the time series to continue, we expect the forecasted values for the future time periods to be relatively accurate.

We can also use the measures of forecast accuracy to compare different forecasts for the same time series, which will allow us to pick the best forecast for that time series. For example, we can compare the *MADs* for different forecasts and then pick the forecast with the best (i.e. smallest) *MAD*. Similarly, we can compare the *MSEs* or *MAPEs* for different forecasts. When comparing measures of forecast accuracy for different forecasts, we have to compare the same measure for each forecast to get a meaningful comparison. We cannot compare the *MAD* from one forecast with the *MSE* from another forecast because the *MAD* and *MSE* are analyzing the errors in different ways.

EXAMPLE

A local company produces and sells a certain product. The number of units sold each month for a year is recorded in the table below. The company created two forecasts to predict the number of units sold each month. Which forecast should the company use?

Month	Actual Number of Units Sold	Forecast 1	Forecast 2
January	83	82	75
February	102	105	105
March	75	82	83
April	125	117	120
May	100	95	107
June	103	111	99
July	118	115	113
August	101	97	104
September	70	78	75
October	99	90	97
November	120	125	123
December	97	103	100

Solution

The MAD for Forecast 1 is 5.583. The MAD for Forecast 2 is 4.667. Because the MAD for Forecast 2 is smaller, it suggests that Forecast 2 is more accurate. So, the company should use Forecast 2.

NOTES

1. We could also compare the *MSE*'s or *MAPE*'s for these forecasts. The *MSE* for Forecast 1 is **36.917** and the *MSE* for Forecast 2 is **25.667**. So, based on the *MSE*'s, the company should use Forecast 2. The *MAPE* for Forecast 1 is 5.84\% and the *MAPE* for Forecast 2 is 5.01\%. So, based on the *MAPE*'s, the company should use Forecast 2.
2. In this example, the three measures of forecast accuracy all agree that Forecast 2 is the most accurate. But it is possible that the measures of forecast accuracy do not agree on which forecast is most accurate. That is, one measure of forecast accuracy might indicate one forecast is more accurate, but another measure of forecast accuracy might indicate a different forecast is most accurate. We do not need to compare all three measures of forecast accuracy. In general, pick one of the measures and use that measure to assess the different forecasts.

When assessing different forecasts for the same time series, comparing the measures of forecast accuracy is an important tool to assess how well the forecast fits the data. But the measures of forecast accuracy should not be the only factor we consider when assessing a forecast. We also need to rely on our own judgement and consider other factors, such as current business or economic conditions, that might affect the forecast.

TRY IT

Two different forecasts were created for the following time series.

Period	Actual Value	Forecast 1	Forecast 2
1	52	50	
2	48	55	47
3	51	58	44
4	51	46	54
5	60	55	55
6	59	50	53
7	40	47	49
8	45	48	52

1. Calculate the MAD for each forecast.
2. Based on the MAD s, which forecast is most accurate?
3. Calculate the MSE for each forecast.
4. Based on the MSE s, which forecast is most accurate?
5. Calculate the $MAPE$ for each forecast.
6. Based on the $MAPE$ s, which forecast is most accurate?

Click to see Solution

1.	Period	Actual Value	Forecast 1	Error (Forecast 1)	Forecast 2	Error (Forecast 2)
	1	52	50	2		
	2	48	55	7	47	1
	3	51	58	7	44	7
	4	51	46	5	54	3
	5	60	55	5	55	5
	6	59	50	9	53	6
	7	40	47	7	49	9
	8	45	48	3	52	7
			Sum	45	Sum	38

$$\begin{aligned}
 MAD \text{ for Forecast 1} &= \frac{\sum |\text{forecast error}|}{\text{number of forecast errors}} \\
 &= \frac{45}{8} \\
 &= 5.625
 \end{aligned}$$

$$\begin{aligned}
 MAD \text{ for Forecast 2} &= \frac{\sum |\text{forecast error}|}{\text{number of forecast errors}} \\
 &= \frac{38}{7} \\
 &= 5.429
 \end{aligned}$$

2. Forecast 2 because it has the smaller *MAD*.

3.

Period	Actual Value	Forecast 1	(Error) ² (Forecast 1)	Forecast 2	(Error) ² (Forecast 2)
1	52	50	4		
2	48	55	49	47	1
3	51	58	49	44	49
4	51	46	25	54	9
5	60	55	25	55	25
6	59	50	81	53	36
7	40	47	49	49	81
8	45	48	9	52	49
		Sum	291	Sum	250

$$\begin{aligned}
 MSE \text{ for Forecast 1} &= \frac{\sum(\text{forecast error})^2}{\text{number of forecast errors}} \\
 &= \frac{291}{8} \\
 &= 36.375
 \end{aligned}$$

$$\begin{aligned}
 MSE \text{ for Forecast 2} &= \frac{\sum(\text{forecast error})^2}{\text{number of forecast errors}} \\
 &= \frac{250}{7} \\
 &= 35.714
 \end{aligned}$$

4. Forecast 2 because it has the smaller *MSE*.

5.

Period	Actual Value	Forecast 1	Percent Error (Forecast 1)	Forecast 2	Percent Error (Forecast 2)
1	52	50	0.0384...		
2	48	55	0.1458...	47	0.0208...
3	51	58	0.1372...	44	0.1372...
4	51	46	0.0980...	54	0.0588...
5	60	55	0.0833...	55	0.0833...
6	59	50	0.1525...	53	0.1016...
7	40	47	0.175	49	0.225
8	45	48	0.0666...	52	0.1555...
		Sum	0.8971...	Sum	0.7824...

$$\begin{aligned}
 \text{MAPE for Forecast 1} &= \frac{\sum |\text{percent error}|}{\text{number of forecast errors}} \times 100\% \\
 &= \frac{0.8971}{8} \times 100\% \\
 &= 11.21\% \\
 \text{MAPE for Forecast 2} &= \frac{\sum |\text{percent error}|}{\text{number of forecast errors}} \times 100\% \\
 &= \frac{0.7824}{7} \times 100\% \\
 &= 11.18\%
 \end{aligned}$$

6. Forecast 2 because it has the smaller *MAPE*.

Exercises

1. The number of pizzas sold each week at a local pizza restaurant, along with two different forecasts, are given in the table below.

Week	Number of Pizzas Sold	Forecast 1	Forecast 2
1	120	123	
2	135	126	
3	140	129	128
4	115	132	138
5	130	135	128
6	145	139	123
7	160	142	138
8	170	145	153
9	125	148	165
10	130	151	148
11	150	155	128
12	140	158	140
13	155	161	144
14	180	164	148
15	165	167	168
16	170	171	173
17	190	174	168

- a. Calculate the MAD , MSE , and $MAPE$ for Forecast 1.
- b. Interpret the MAD for Forecast 1.
- c. Calculate the MAD , MSE , and $MAPE$ for Forecast 2.
- d. Interpret the $MAPE$ for Forecast 2.
- e. Based on the MAD , which forecast is more accurate? Why?

Click to see Answer

- a. $MAD = 11.88$, $MSE = 198.94$, $MAPE=8.20\%$
- b. On average, the forecasted values differ by 11.88 pizzas sold from the actual values.
- c. $MAD = 16.53$, $MSE = 397.87$, $MAPE=11.26\%$

- d. On average, the forecasted values differ by 11.26\% from the actual values.
 - e. Forecast 1 because it has the smaller MAD .
2. The number of coffees sold each day at a local coffee shop, along with two different forecasts, are given in the table below.

Day	Number of Coffees Sold	Forecast 1	Forecast 2
1	120		
2	100		
3	115		
4	130		111
5	110	116	123
6	150	114	117
7	165	126	136
8	110	139	155
9	120	134	131
10	125	136	122
11	135	130	122
12	130	123	131
13	152	128	131
14	160	136	144
15	110	144	155
16	120	138	129
17	125	136	121
18	127	129	122
19	130	121	126
20	155	126	129
21	162	134	145
22	120	144	157
23	125	142	136
24	130	141	127
25	135	134	128
26	137	128	133
27	160	132	136
28	155	141	151
29	130	147	155

30	140	146	141
31	145	146	139
32	139	143	142
33	142	139	141
34	165	142	141
35	163	148	156
36	130	152	162
37	120	150	143
38	125	145	127
39	135	135	124
40	145	128	131
41	160	131	140
42	170	141	153
43	130	153	165
44	137	151	145
45	130	149	138
46	135	142	132
47	132	133	134
48	163	134	133
49	168	140	151
50	125	150	163
51	130	147	142
52	120	147	132
53	115	136	124
54	125	123	118
55	156	123	122
56	161	129	143

- Calculate the MAD , MSE , and $MAPE$ for Forecast 1.
- Calculate the MAD , MSE , and $MAPE$ for Forecast 2.
- Based on the MSE , which forecast is more accurate? Why?

Click to see Answer

- a. $MAD = 17.77$, $MSE = 431.12$, $MAPE = 12.79\%$
- b. $MAD = 15.53$, $MSE = 382.23$, $MAPE = 11.41\%$
- c. Forecast 2 because it has the smaller MSE .

3. The quarterly demand, in 1000s, for a particular product, along with two different forecasts, are recorded in the table below.

Quarter	Demand (in 1000s)	Forecast 1	Forecast 2
1	50	32	
2	75	80	
3	85	88	
4	70	72	70
5	60	51	77
6	95	99	72
7	100	107	75
8	85	91	85
9	70	70	93
10	115	118	85
11	125	126	90
12	110	110	103
13	80	89	117
14	140	137	105
15	145	145	110
16	130	129	122
17	90	108	138
18	165	156	122
19	175	164	128
20	155	148	143

- a. Calculate the MAD , MSE , and $MAPE$ for Forecast 1.

- b. Interpret the $MAPE$ for Forecast 1.
- c. Calculate the MAD , MSE , and $MAPE$ for Forecast 2.
- d. Interpret the MAD for Forecast 2.
- e. Based on the $MAPE$, which forecast is more accurate? Why?

Click to see Answer

- a. $MAD = 5.8$, $MSE = 61$, $MAPE=6.81\%$
- b. On average, the forecasted values differ by 6.81\% from the actual values.
- c. $MAD = 25$, $MSE = 855$, $MAPE=22.73\%$
- d. On average, the forecasted values differ by 25, 000 from the actual values.
- e. Forecast 1 because it has the smaller $MAPE$.

14.3 SMOOTHING MODELS

LEARNING OBJECTIVES

- Construct forecast models using moving averages, weighted moving averages, and exponential smoothing.

Although some time series do not contain trend, seasonal, or cyclical patterns, almost every time series contains some irregular patterns or random variations. In this section, we will discuss three models—moving averages, weighted moving averages, and exponential smoothing—to reduce or “smooth out” the irregular patterns in the time series. Because the purpose of these models is to “smooth out” the random variations, these models are called **smoothing models**.

The moving averages, weighted moving averages, and exponential smoothing models are only good at “smoothing” out the random variations present in a time series. When random variation is the only component present in the time series, these models generally create accurate forecasts. But these three models have their limitations. For example, these three models can only make a predication about the next time period in the time series, and so they are not appropriate for more long-term forecasts. Also, these smoothing models are not good at modelling any significant trend, seasonal, or cyclical patterns present in the time series. In cases where trend, seasonality, or cycles are present in the time series, a model that deals with those types of patterns should be used to create an accurate forecast.

The three models discussed in this section are called **averaging models**. In slightly different ways, each model “averages” data from several previous time periods to create a forecast for the next time period. By using an averaging technique, this type of model smooths out extremes or irregularities in the data.

Moving Averages

The **moving average model** averages the most recent k values in the time series to forecast the next time period.

forecast for each time period = average of the previous k observations

The “moving” part of this model comes from the fact that as new data becomes available, the oldest observation is replaced with the newest observation, and the average is recalculated. So, the average “moves” or “changes” as each new observation becomes available.

Any number of time periods may be used to calculate a moving average forecast, but the number of time periods used in the forecast must be consistent across the model. That is, we cannot do a 3-month moving average for part of the forecast and then switch to a 4-month moving average for the rest of the forecast. In general, the forecast becomes smoother when more time periods are included in the average. But too much smoothing may suppress other components of the time series. Through the use of technology, such as Excel, it is easy to experiment with different numbers of time periods in a moving average forecast to find the best forecast.

Suppose we want to create a 2-month moving average forecast. The first forecast is for the third month and is the average of the time series values from months one and two. The next forecast is for the fourth month and is the average of the time series values from months two and three. In general, the forecast for each month is the average of the values from the previous two months.

EXAMPLE

A local company produces and sells a certain product. The number of units sold each month for a year is recorded in the table below.

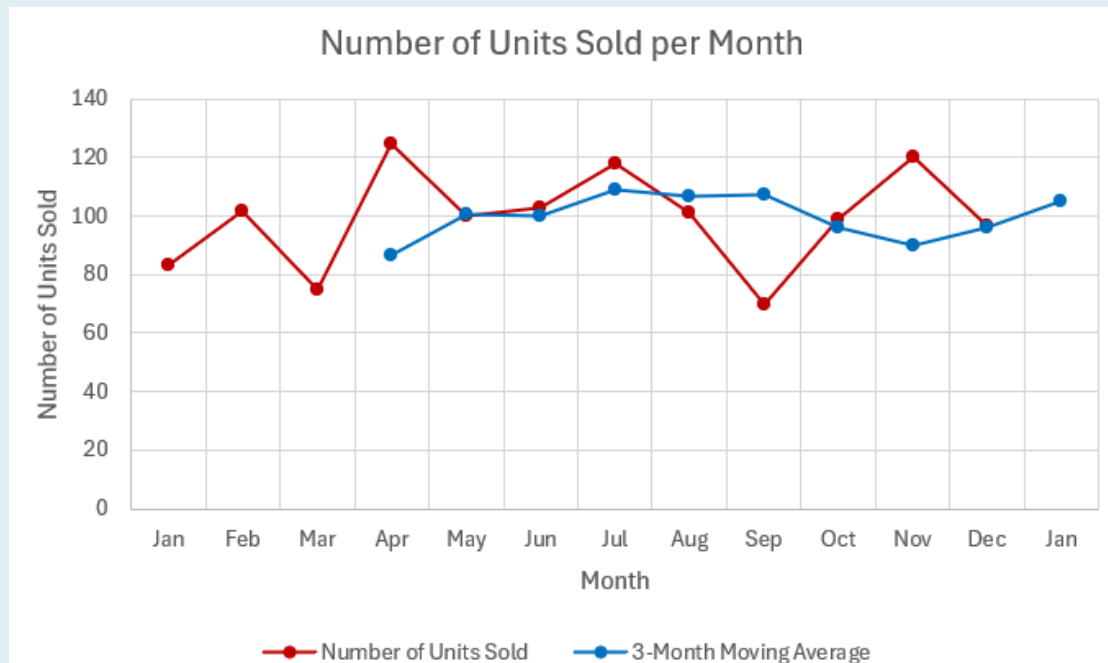
Month	Actual Number of Units Sold
January	83
February	102
March	75
April	125
May	100
June	103
July	118
August	101
September	70
October	99
November	120
December	97

1. Create a 3-month moving average forecast for this time series.
2. What is the forecast for January of the next year?
3. Calculate the *MAD* for the 3-month moving average forecast.

Solution

1.

Month	Actual Number of Units Sold	3-Month Moving Average Forecast
January	83	
February	102	
March	75	
April	125	86.666...
May	100	100.666...
June	103	100
July	118	109.333...
August	101	107
September	70	107.333...
October	99	96.333...
November	120	90
December	97	96.333...
January		105.333...



2. The forecast for January of the next year is 105.333...

3.

Month	Actual Number of Units Sold	3-Month Moving Average Forecast	Forecast Error
January	83		
February	102		
March	75		
April	125	86.666 ...	38.333 ...
May	100	100.666 ...	0.666 ...
June	103	100	3
July	118	109.333 ...	8.666 ...
August	101	107	6
September	70	107.333 ...	37.333 ...
October	99	96.333 ...	2.666 ...
November	120	90	30
December	97	96.333 ...	0.666 ...
January		105.333 ...	
		Sum	127.333 ...

$$\begin{aligned}
 MAD &= \frac{\sum |\text{forecast error}|}{\text{number of forecast errors}} \\
 &= \frac{127.333 \dots}{9} \\
 &= 14.15
 \end{aligned}$$

NOTES

1. To calculate the forecast for April, we average the values from the previous three months:

$$\text{Forecast for April} = \frac{83 + 102 + 75}{3} = 86.666 \dots$$

To calculate the forecast for May, we average the values from the previous three months:

$$\text{Forecast for May} = \frac{102 + 75 + 125}{3} = 100.666 \dots$$

This process continues for each month, averaging the previous three months of observations.

2. The graph shows the actual time series data and the **3**-month moving average forecast. Note how the **3**-month moving average has smoothed out the monthly variations and indicates the overall trend of the monthly units sold.
3. There is no forecast for the first three months. In order to construct a **3**-month moving average forecast, we need to have three months of prior observations to create the forecast. For January, February, and March, we do not have three months of previous observations, so we cannot create a forecast for those months.
4. January of the next year is the last time period that has a forecast because that is the last time period where we have three months of prior observations. We cannot create a forecast for February of the next year because there is no actual value for January of the next year, and so we do not have three months of previous observations.
5. In the calculation of the *MAD*, remember that we can only calculate errors when there is an actual value **and** a forecast value. In this example, there are no errors for January, February, and March because there are no forecasts for those months. Also, there is no error for January of the next year because there is no actual value for that month. Altogether, there are only nine errors, and the *MAD* is the average of the absolute value of those nine errors.

TRY IT

A local company produces and sells a certain product. The number of units sold each month for a year is recorded in the table below.

Month	Actual Number of Units Sold
January	83
February	102
March	75
April	125
May	100
June	103
July	118
August	101
September	70
October	99
November	120
December	97

1. Create a 4-month moving average forecast for this time series.
2. What is the forecast for January of the next year?
3. Calculate the *MAD* for the 4-month moving average forecast.
4. Which forecast is better: the 3-month moving average (from the example above) or the 4-month moving average? Why?

Click to see Solution

1.

Month	Actual Number of Units Sold	4-Month Moving Average Forecast	Forecast Error
January	83		
February	102		
March	75		
April	125		
May	100	96.25	3.75
June	103	100.5	2.5
July	118	100.75	17.25
August	101	111.75	10.5
September	70	105.5	35.5
October	99	98	1
November	120	97	23
December	97	97.5	0.5
January		96.5	
		Sum	94

2. The forecast for January of the next year is 96.5.

3. $MAD = \frac{\sum |\text{forecast error}|}{\text{number of forecast errors}} = \frac{94}{8} = 11.75$

4. The 4-month moving average is better because it has the smaller *MAD*.

Weighted Moving Averages

In the moving average model, each observation in the average calculation receives the same weight. But what if we want to apply more emphasis or weight to certain observations and less weight to others? The **weighted moving average model** applies different pre-set weight to each of the previous values and then calculates the weighted average of the most recent k values in the time series to forecast the next time period.

$$\text{forecast for each time period} = \sum [(\text{weight } i) \times (\text{previous value } i)]$$

To calculate the weighted moving average, multiply each observation by its corresponding weight

and then add up the results. Typically, more recent observations receive higher weights, giving greater emphasis to more recent values in the time series, and older observations receive lower weights, giving less emphasis to older values in the time series.

For a k -period weighted moving average, the k pre-set weights are applied to the previous k observations. The weights used in a weighted moving average forecast are numbers between 0 and 1, and the sum of the weights must equal 1.

Any number of time periods may be used to calculate a weighted moving average forecast, but the number of time periods used in the forecast must be consistent across the model. The weights must be determined before creating the forecast, and those predetermined weights must be consistent across the entire forecast. Different weights will produce different forecasts. Through the use of technology, such as Excel, it is easy to experiment with different weights and different numbers of time periods in a weighted moving average forecast to find the best forecast.

EXAMPLE

A local company produces and sells a certain product. The number of units sold each month for a year is recorded in the table below.

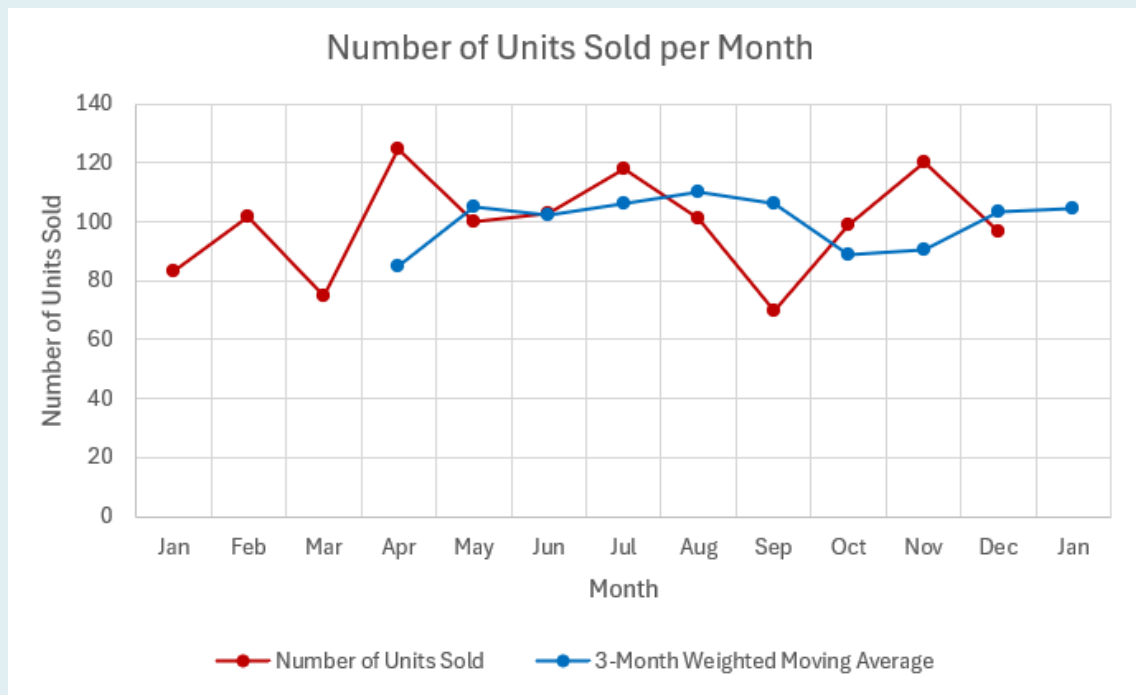
Month	Actual Number of Units Sold
January	83
February	102
March	75
April	125
May	100
June	103
July	118
August	101
September	70
October	99
November	120
December	97

1. Create a 3-month weighted moving average forecast for this time series. Assign a weight of 0.5 to the most recent observation, a weight of 0.3 to the second most recent observation, and a weight of 0.2 to the third most recent observation.
2. What is the forecast for January of the next year?
3. Calculate the MSE for the 3-month weighted moving average forecast.

Solution

1.

Month	Actual Number of Units Sold	3-Month Weighted Moving Average Forecast
January	83	
February	102	
March	75	
April	125	84.7
May	100	105.4
June	103	102.5
July	118	106.5
August	101	109.5
September	70	106.5
October	99	88.9
November	120	90.7
December	97	103.7
January		104.3



2. The forecast for January of the next year is 104.3.

3.

Month	Actual Number of Units Sold	3-Month Weighted Moving Average Forecast	(Forecast Error) ²
January	83		
February	102		
March	75		
April	125	84.7	1624.09
May	100	105.4	29.16
June	103	102.5	0.25
July	118	106.5	132.25
August	101	109.5	79.21
September	70	106.5	1332.25
October	99	88.9	102.01
November	120	90.7	858.49
December	97	103.7	44.89
January		104.3	
		Sum	4202.6

$$\begin{aligned}
 MSE &= \frac{\sum(\text{forecast error})^2}{\text{number of forecast errors}} \\
 &= \frac{4202.6}{9} \\
 &= 466.96
 \end{aligned}$$

NOTES

1. To calculate the forecast for April, we multiply each of the previous three months by their corresponding weights and add up the results:

$$\text{Forecast for April} = 0.5 \times 75 + 0.3 \times 102 + 0.2 \times 83 = 84.7$$

To calculate the forecast for May, we multiply each of the previous three months by their corresponding weights and add up the results:

$$\text{Forecast for May} = 0.5 \times 125 + 0.3 \times 75 + 0.2 \times 102 = 105.4$$

This process continues for each month, multiplying each of the previous three months by their corresponding weights and adding up the results.

2. The graph shows the actual time series data and the 3-month weighted moving average forecast. Note how the 3-month weighted moving average has smoothed out the monthly variations and indicates the overall trend of the monthly units sold.
3. There is no forecast for the first three months. In order to construct a 3-month weighted moving average forecast, we need to have three months of prior observations to create the forecast. For January, February, and March, we do not have three months of previous observations, so we cannot create a forecast for those months.
4. January of the next year is the last time period that has a forecast because that is the last time period where we have three months of prior observations. We cannot create a forecast for February of the next year because there is no actual value for January of the next year, and so we do not have three months of previous observations.
5. In the calculation of the *MSE*, remember that we can only calculate errors when there is an actual value **and** a forecast value. In this example, there are no errors for January, February, and March because there are no forecasts for those months. Also, there is no error for January of the next year because there is no actual value for that month. Altogether, there are only nine errors, and the *MSE* is the average of the squares of those nine errors.

TRY IT

A local company produces and sells a certain product. The number of units sold each month for a year is recorded in the table below.

Month	Actual Number of Units Sold
January	83
February	102
March	75
April	125
May	100
June	103
July	118
August	101
September	70
October	99
November	120
December	97

1. Create a 5-month weighted moving average forecast for this time series. Assign a weight of 0.4 to the most recent observation, a weight of 0.2 to the second and third most recent observation, and a weight of 0.1 to the fourth and fifth most recent observation.
2. What is the forecast for January of the next year?
3. Calculate the MSE for the 5-month weighted moving average forecast.
4. Which forecast is better: the 3-month weighted moving average (from the example above) or the 5-month weighted moving average? Why?

Click to see Solution

1.

Month	Actual Number of Units Sold	5-Month Weighted Moving Average Forecast	(Forecast Error) ²
January	83		
February	102		
March	75		
April	125		
May	100		
June	103	98.5	20.25
July	118	103.9	198.81
August	101	107.8	46.24
September	70	107.1	1376.41
October	99	92.1	47.61
November	120	95.9	580.81
December	97	103.7	44.89
January		99.7	
		Sum	2315.02

2. The forecast for January of the next year is 99.7.

3.
$$MSE = \frac{\sum(\text{forecast error})^2}{\text{number of forecast errors}} = \frac{2315.02}{7} = 330.72$$

4. The 5-month weighted moving average is better because it has the smaller MSE .

Exponential Smoothing

The **exponential smoothing model** is a special case of a weighted moving average model. In an exponential smoothing model there is only one weight—the weight for the most recent observation. The weight given to the most recent observation, denoted by α , is called the **smoothing constant**. The remaining weight, $1 - \alpha$, is applied to the previous forecast.

$$\text{forecast for each time period} = \alpha \times (\text{previous value}) + (1 - \alpha) \times (\text{previous forecast})$$

In exponential smoothing, the weighting on the other values decrease exponentially as the observations move further away from the forecasted time period. These other weights are calculated

automatically by the formula—all we need to produce an exponential smoothing forecast is the value of smoothing constant α .

The smoothing constant α is a pre-set weight and is a number between 0 and 1. A larger value of α places more emphasis on the most recent observation, and a smaller value of α places less emphasis on the most recent observation. The smoothing constant must be determined before creating the forecast, and the value of the smoothing constant must be consistent across the entire forecast. Different values of the smoothing constant will produce different forecasts. Through the use of technology, such as Excel, it is easy to experiment with different values of the smoothing constant in an exponential smoothing forecast to find the best forecast.

NOTE

To create an exponential smoothing forecast, an initial forecast value is needed for the first time period. In some instances, an initial forecast value may be provided. If no initial forecast value is given, use the actual time series value from the first time period as the initial forecast.

EXAMPLE

A local company produces and sells a certain product. The number of units sold each month for a year is recorded in the table below.

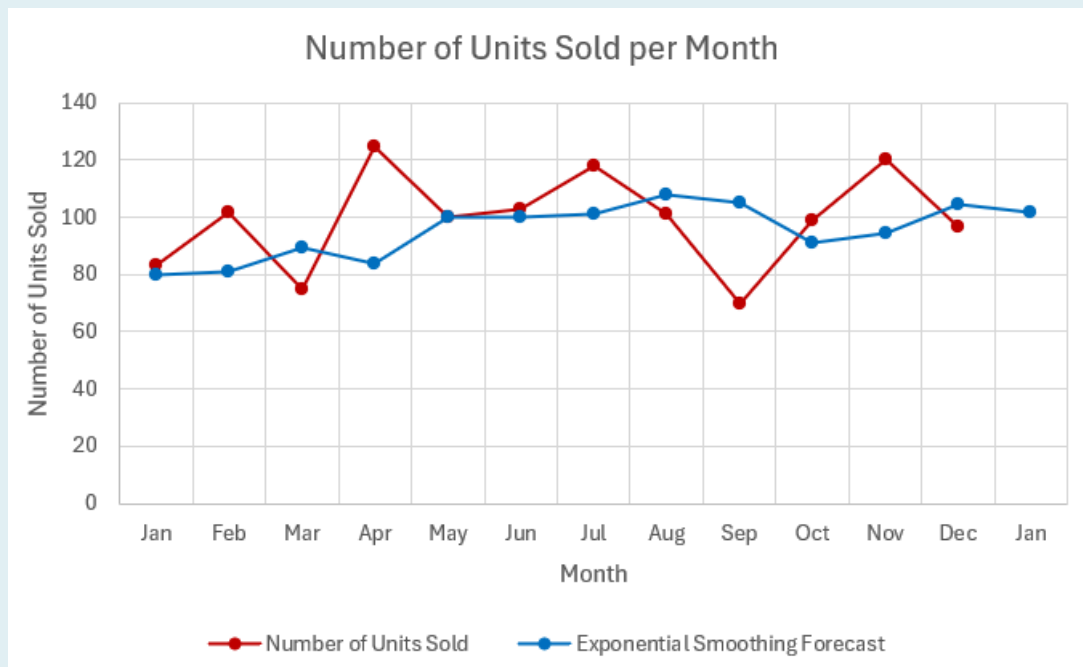
Month	Actual Number of Units Sold
January	83
February	102
March	75
April	125
May	100
June	103
July	118
August	101
September	70
October	99
November	120
December	97

1. Create an exponential smoothing forecast for this time series. Use a smoothing constant of 0.4 and an initial forecast of 80.
2. What is the forecast for January of the next year?
3. Calculate the *MAPE* for the exponential smoothing forecast.

Solution

1.

Month	Actual Number of Units Sold	Exponential Smoothing Forecast
January	83	80
February	102	81.2
March	75	89.52
April	125	83.712
May	100	100.227 ...
June	103	100.136 ...
July	118	101.281 ...
August	101	107.969 ...
September	70	105.181 ...
October	99	91.108 ...
November	120	94.265 ...
December	97	104.559 ...
January		101.535 ...



2. The forecast for January of the next year is 101.535 ...

3.

Month	Actual Number of Units Sold	Exponential Smoothing Forecast	Percent Error
January	83	80	0.0361 ...
February	102	81.2	0.2039 ...
March	75	89.52	0.1936
April	125	83.712	0.3303 ...
May	100	100.227 ...	0.0022 ...
June	103	100.136 ...	0.0278 ...
July	118	101.281 ...	0.1416 ...
August	101	107.969 ...	0.0690 ...
September	70	105.181 ...	0.5025 ...
October	99	91.108 ...	0.0797 ...
November	120	94.265 ...	0.2144 ...
December	97	104.559 ...	0.0779 ...
January		101.535 ...	
		Sum	1.879 ...

$$\begin{aligned} \text{MAPE} &= \frac{\sum |\text{percent error}|}{\text{number of forecast errors}} \times 100\% \\ &= \frac{1.879}{12} \times 100\% \\ &= 15.66\% \end{aligned}$$

NOTES

1. The initial forecast for January is **80**, as provided in the question. If no initial forecast is provided, use the initial actual value as the first forecast value.
2. To calculate the forecast for February, multiply the actual value for January by the smoothing constant **0.4**, multiply the forecast for January by $1 - 0.4 = 0.6$, and then add up the results:

$$\text{Forecast for February} = 0.4 \times 83 + 0.6 \times 80 = 81.2$$

To calculate the forecast for March, multiply the actual value for February by the smoothing

constant 0.4 , multiply the forecast for February by $1 - 0.4 = 0.6$, and then add up the results:

$$\text{Forecast for March} = 0.4 \times 102 + 0.6 \times 81.2 = 89.52$$

This process continues for each month, multiplying each previous value by the smoothing constant 0.4 , multiplying each previous forecast by $1 - 0.4 = 0.6$, and then adding up the results.

3. The graph shows the actual time series data and the exponential smoothing forecast. Note how the exponential smoothing forecast has smoothed out the monthly variations and indicates the overall trend of the monthly units sold.
4. January of the next year is the last time period that has a forecast because that is the last time period where we have both a prior actual value and a prior forecast. We cannot create a forecast for February of the next year because there is no actual value for January of the next year.
5. In the calculation of the *MAPE*, remember that we can only calculate errors when there is an actual value **and** a forecast value. In this example, there is no error for January of the next year because there is no actual value for that month. Altogether, there are only twelve errors, and the *MAPE* is the average of the absolute value of the percent errors of those twelve errors.

TRY IT

A local company produces and sells a certain product. The number of units sold each month for a year is recorded in the table below.

Month	Actual Number of Units Sold
January	83
February	102
March	75
April	125
May	100
June	103
July	118
August	101
September	70
October	99
November	120
December	97

1. Create an exponential smoothing forecast with $\alpha = 0.7$ for this time series.
2. What is the forecast for January of the next year?
3. Calculate the *MAPE* for the exponential smoothing forecast.
4. Which forecast is better: the exponential smoothing with $\alpha = 0.4$ (from the example above) or the exponential smoothing with $\alpha = 0.7$? Why?

Click to see Solution

1.

Month	Actual Number of Units Sold	Exponential Smoothing Forecast	Percent Error
January	83	83	0
February	102	83	0.186
March	75	96.3	0.284
April	125	81.39	0.348...
May	100	111.917	0.119...
June	103	103.5751	0.005...
July	118	103.172...	0.125...
August	101	113.551...	0.124...
September	70	104.765...	0.496...
October	99	80.429...	0.187...
November	120	93.428...	0.221...
December	97	112.028...	0.154...
January		101.508...	
		Sum	2.254...

2. The forecast for January of the next year is 101.508...
3.
$$\text{MAPE} = \frac{\sum |\text{percent error}|}{\text{number of forecast errors}} \times 100\% = \frac{2.254}{12} \times 100\% = 18.79\%$$
4. The exponential smoothing forecast with $\alpha = 0.4$ is better because it has the smaller *MAPE*.

Exercises

1. Consider the time series given in the table below.

Month	Value
1	20
2	25
3	11
4	16
5	19
6	25
7	22
8	14
9	17

- Create a 2-month moving average forecast for this time series. What is the forecast for month 10? Calculate the *MAD* for this forecast.
- Create a 3-month weighted moving average forecast for this time series with weights of 0.6 for the most recent observation, 0.3 for the second most recent observation, and 0.1 for the third most recent observation. What is the forecast for month 10? Calculate the *MAD* for this forecast.
- Create an exponential smoothing forecast for this time series with a smoothing constant of 0.8. What is the forecast for month 10? Calculate the *MAD* for this forecast.
- Of the three forecasts created above, which appears to provide the most accurate forecast based on *MAD*? Explain.

Click to see Answer

- forecast for month 10 = 15.5, *MAD* = 5.29

Month	Value	Forecast
1	20	
2	25	
3	11	22.5
4	16	18
5	19	13.5
6	25	17.5
7	22	22
8	14	23.5
9	17	18
10		15.5

b. forecast for month 10 = 16.6, $MAD = 3.47$

Month	Value	Forecast
1	20	
2	25	
3	11	
4	16	16.1
5	19	15.4
6	25	17.3
7	22	22.3
8	14	22.6
9	17	17.5
10		16.6

c. forecast for month 10 = 16.73, $MAD = 4.66$

Month	Value	Forecast
1	20	20
2	25	20
3	11	24
4	16	13.6
5	19	15.52
6	25	13.304
7	22	23.6608
8	14	22.33216
9	17	15.666432
10		16.7332864

d. The 3-month weighted moving average is the most accurate forecast because it has the smallest *MAD*.

2. The number of pizzas sold each week at a local pizza restaurant is recorded in the table below.

Week	Number of Pizzas Sold
1	120
2	135
3	140
4	115
5	130
6	145
7	160
8	170
9	125
10	130
11	150
12	140
13	155
14	180
15	165
16	170
17	190

- Create a 4-week moving average forecast for this time series. What is the forecast for week 18? Calculate the $MAPE$ for this forecast.
- Create a 4-week weight moving average forecast for this time series with weights of 0.5 for the most recent observation, 0.3 for the second most recent observation, and 0.1 for the third and fourth most recent observation. What is the forecast for week 18? Calculate the $MAPE$ for this forecast.
- Create an exponential smoothing forecast for this time series with a smoothing constant of 0.6 and an initial forecast 125. What is the forecast for week 18? Calculate the $MAPE$ for this forecast.
- Based on the $MAPE$, which forecast is more accurate? Why?

Click to see Answer

- forecast for week 18 = 176.25, MAPE=11.18\%

Week	Number of Pizzas Sold	Forecast
1	120	
2	135	
3	140	
4	115	
5	130	127.5
6	145	130
7	160	132.5
8	170	137.5
9	125	151.25
10	130	150
11	150	146.25
12	140	143.75
13	155	136.25
14	180	143.75
15	165	156.25
16	170	160
17	190	167.5
18		176.25

- b. forecast for week 18 = 180.5, MAPE=10.5\%

Week	Number of Pizzas Sold	Forecast
1	120	
2	135	
3	140	
4	115	
5	130	125
6	145	127
7	160	137
8	170	148
9	125	160.5
10	130	144
11	150	135.5
12	140	143.5
13	155	140.5
14	180	147.5
15	165	165.5
16	170	166
17	190	168
18		180.5

c. forecast for week 18 = 181.38, MAPE=10.19\%

Week	Number of Pizzas Sold	Forecast
1	120	125
2	135	122
3	140	129.8
4	115	135.92
5	130	123.368
6	145	127.347 ...
7	160	137.938 ...
8	170	151.175 ...
9	125	162.470 ...
10	130	139.988 ...
11	150	133.995 ...
12	140	143.598 ...
13	155	141.439 ...
14	180	149.575 ...
15	165	167.830 ...
16	170	166.132 ...
17	190	168.452 ...
18		181.381 ...

d. Exponential smoothing because it has the smallest *MAPE*.

3. The number of coffees sold each day at a local coffee shop is given in the table below.

Day	Number of Coffees Sold
1	120
2	100
3	115
4	130
5	110
6	150
7	165
8	110
9	120
10	125
11	135
12	130
13	152
14	160
15	110
16	120
17	125
18	127
19	130
20	155
21	162
22	120
23	125
24	130
25	135
26	137
27	160
28	155
29	130

30	140
31	145
32	139
33	142
34	165
35	163
36	130
37	120
38	125
39	135
40	145
41	160
42	170
43	130
44	137
45	130
46	135
47	132
48	163
49	168
50	125
51	130
52	120
53	115
54	125
55	156
56	161

- Create an exponential smoothing forecast with a smoothing constant of 0.3. Calculate the MAD for this forecast.
- Create an exponential smoothing forecast with a smoothing constant of 0.9. Calculate

the MAD for this forecast.

- c. Based on the MAD , which forecast is more accurate? Why?

Click to see Answer

- a. $MAD = 14.79$

Day	Number of Coffees Sold	Forecast
1	120	120
2	100	120
3	115	114
4	130	114.3
5	110	119.01
6	150	116.307
7	165	126.4149
8	110	137.99043
9	120	129.593...
10	125	126.715...
11	135	126.200...
12	130	128.840...
13	152	129.188...
14	160	136.031...
15	110	143.222...
16	120	133.255...
17	125	129.278...
18	127	127.995...
19	130	127.696...
20	155	128.387...
21	162	136.371...
22	120	144.059...
23	125	136.841...
24	130	133.289...
25	135	132.302...
26	137	133.111...
27	160	134.278...
28	155	141.994...
29	130	145.896...

30	140	141.127...
31	145	140.789...
32	139	142.052...
33	142	141.136...
34	165	141.395...
35	163	148.476...
36	130	152.833...
37	120	145.983...
38	125	138.188...
39	135	134.232...
40	145	134.462...
41	160	137.623...
42	170	144.336...
43	130	152.035...
44	137	145.424...
45	130	142.897...
46	135	139.028...
47	132	137.819...
48	163	136.073...
49	168	144.151...
50	125	151.306...
51	130	143.414...
52	120	139.390...
53	115	133.573...
54	125	128.001...
55	156	127.100...
56	161	135.770...
57		143.339...

b. $MAD = 14.06$

Day	Number of Coffees Sold	Forecast
1	120	120
2	100	120
3	115	102
4	130	113.7
5	110	128.37
6	150	111.837
7	165	146.1837
8	110	163.118...
9	120	115.311...
10	125	119.531...
11	135	124.453...
12	130	133.945...
13	152	130.394...
14	160	149.839...
15	110	158.983...
16	120	114.898...
17	125	119.489...
18	127	124.448...
19	130	126.744...
20	155	129.674...
21	162	152.467...
22	120	161.046...
23	125	124.104...
24	130	124.910...
25	135	129.491...
26	137	134.449...
27	160	136.744...
28	155	157.674...
29	130	155.267...

30	140	132.526...
31	145	139.252...
32	139	144.425...
33	142	139.542...
34	165	141.754...
35	163	162.675...
36	130	162.967...
37	120	133.296...
38	125	121.329...
39	135	124.632...
40	145	133.963...
41	160	143.896...
42	170	158.389...
43	130	168.838...
44	137	133.883...
45	130	136.688...
46	135	130.668...
47	132	134.566...
48	163	132.256...
49	168	159.925...
50	125	167.192...
51	130	129.219...
52	120	129.921...
53	115	120.992...
54	125	115.599...
55	156	124.059...
56	161	152.805...
57		160.180...

c. A smoothing constant of 0.9 because it has the smaller *MAD*.

4. The quarterly demand, in 1000s, for a particular product is recorded in the table below.

Quarter	Demand (in 1000s)
1	50
2	75
3	85
4	70
5	60
6	95
7	100
8	85
9	70
10	115
11	125
12	110
13	80
14	140
15	145
16	130
17	90
18	165
19	175
20	155

- Create a 5-quarter moving average forecast for this time series. What is the forecast for quarter 21? Calculate the MSE for this forecast.
- Create a 5-quarter weighted moving average forecast for this time series with weights of 0.4 for the most recent observation, 0.3 for the second most recent observation, and 0.1 for the remaining observations. What is the forecast for quarter 21? Calculate the MSE for this forecast.
- Create an exponential smoothing forecast for this time series with a smoothing constant

of 0.4. What is the forecast for quarter 21? Calculate the MSE for this forecast.

d. Based on the MSE , which forecast is more accurate? Why?

Click to see Answer

a. forecast for quarter 21 = 143, $MSE = 792.6$

Quarter	Demand (in 1000s)	Forecast
1	50	
2	75	
3	85	
4	70	
5	60	
6	95	68
7	100	77
8	85	82
9	70	82
10	115	82
11	125	93
12	110	99
13	80	101
14	140	100
15	145	114
16	130	120
17	90	121
18	165	117
19	175	134
20	155	141
21		143

b. forecast for quarter 21 = 153, $MSE = 894.1$

Quarter	Demand (in 1000s)	Forecast
1	50	
2	75	
3	85	
4	70	
5	60	
6	95	66
7	100	79
8	85	90
9	70	86.5
10	115	79
11	125	95
12	110	110
13	80	1018.5
14	140	96
15	145	115
16	130	131.5
17	90	128.5
18	165	111.5
19	175	134.5
20	155	156
21		153

c. forecast for quarter 21 = 152.1, $MSE = 736.14$

Quarter	Demand (in 1000s)	Forecast
1	50	50
2	75	50
3	85	60
4	70	70
5	60	70
6	95	66
7	100	77.6
8	85	86.56
9	70	85.936
10	115	79.561 ...
11	125	93.736 ...
12	110	106.242 ...
13	80	107.745 ...
14	140	96.647 ...
15	145	113.988 ...
16	130	126.392 ...
17	90	127.835 ...
18	165	112.701 ...
19	175	133.620 ...
20	155	150.172 ...
21		152.103 ...

d. Exponential smoothing because it has the smallest MSE .

14.4 SEASONAL INDICES

LEARNING OBJECTIVES

- Calculate and use seasonal indices to create a forecast for a time series.

Seasonal patterns are repeated patterns that occur over a one-year period due to seasonal influences. Repeated and predictable highs and lows in the time series data that occur at the same time each year indicate seasonality in the data. For example, demand for golf clubs or sunscreen lotion are predictably higher during the summer months. Similarly, retail sales experience predictable highs during the holiday season every year. When a seasonal pattern exists, a seasonal index may be used to create a forecast that accounts for the seasonal component in the time series.

A **seasonal index** is a numerical measure of the seasonal variation in the time series. There is a seasonal index for each “season” in the time series. For example, if the time series measures monthly data, each month is a season, and there is a seasonal index for each month. If the time series measures quarterly data, each quarter is a season, and there is a seasonal index for each quarter. A seasonal index represents how much a particular season (i.e. month or quarter) deviates from an average season. An average season has a seasonal index of 1. An above-average season will have a seasonal index greater than 1, and a below-average season will have a seasonal index less than 1.

Calculating Seasonal Indices

The seasonal index for a particular season is found by dividing the average value for that season by the average of all the data.

$$\text{Seasonal Index} = \frac{\text{Average for the Season}}{\text{Average of all Data}}$$

Follow these steps to calculate the seasonal index for a season:

1. Calculate the overall average (mean) of the time series data.
2. Calculate the average (mean) of the data for each season. This is easier to do when the time series data is grouped into seasons (i.e. if the time series measures months, group all the January data points together, group all the February data points together, and so on).
3. Divide the average for each season by the average of all the data. There will be a seasonal index for each season (i.e. if the time series measures months, there will be 12 seasonal indices).

NOTES

1. The seasonal indices described above are typically used on time series that have a seasonal component but no trend component. The seasonal indices described above only capture the seasonal effect in the time series, and so would not create a good forecast when both seasonality and trend are present in the time series.
2. When both seasonality and trend are present in a time series, a change from one season to the next may be due to a trend, a seasonal variation, or just a random fluctuation. In such cases, seasonal indices are calculated using the centred moving average approach. The centred moving average approach to seasonal indices prevents a variation due to a trend from being incorrectly identified as a variation due to seasonality. The seasonal indices based on centred moving averages remove the effect of the season so that the trend effect is easier to identify. The seasonal indices based on centred moving averages are typically used in the decomposition model, which isolates the seasonal and trend components in a time series. Both the seasonal indices based on centred moving averages and the decomposition model are not discussed in this book.

EXAMPLE

Falcon Golf Supply Company sells a wide range of products for golfers. The company recorded two years of monthly sales for its best-selling set of golf clubs, the Peregrine Set.

Month	Year 1	Year 2
January	70	61
February	72	72
March	85	79
April	101	103
May	123	120
June	108	117
July	99	100
August	92	95
September	80	85
October	65	81
November	69	73
December	82	86

1. Calculate the seasonal index for each month.
2. Identify the average, above-average, and below-average seasons.
3. Suppose the company predicts that they will sell a total of 2,400 sets of Peregrine clubs in year 3. Forecast the sales for each month of year 3.

Solution

1. The average of all of the data is 88.25.

Month	Year 1	Year 2	Average per Month	Seasonal Index
January	70	61	65.5	$\frac{65.5}{88.25} = 0.7422 \dots$
February	72	72	72	$\frac{72}{88.25} = 0.8158 \dots$
March	85	79	82	$\frac{82}{88.25} = 0.9291 \dots$
April	101	103	102	$\frac{102}{88.25} = 1.1558 \dots$
May	123	120	121.5	$\frac{121.5}{88.25} = 1.3767 \dots$
June	108	117	112.5	$\frac{112.5}{88.25} = 1.2747 \dots$
July	99	100	99.5	$\frac{99.5}{88.25} = 1.1274 \dots$
August	92	95	93.5	$\frac{93.5}{88.25} = 1.0594 \dots$
September	80	85	82.5	$\frac{82.5}{88.25} = 0.9384 \dots$
October	65	81	73	$\frac{73}{88.25} = 0.8271 \dots$
November	69	73	71	$\frac{71}{88.25} = 0.8045 \dots$
December	82	86	84	$\frac{84}{88.25} = 0.9518 \dots$

2. April, May, June, July, and August are above-average seasons. January, February, March, September, October, November, and December are below-average seasons.
3. The forecast for all of year 3 is 2,400 sets of clubs, which is $\frac{2,400}{12} = 200$ sets of clubs per month. The forecast of 200 per month is not adjusted for the season. To create a forecast adjusted for the season, we multiply the unadjusted forecast of 200 by the seasonal index for each month.

Month	Seasonal Index	Forecast for Year 3
January	0.7422 ...	$0.7422 \dots \times 200 = 148.44$
February	0.8158 ...	$0.8158 \dots \times 200 = 163.17$
March	0.9291 ...	$0.9291 \dots \times 200 = 185.84$
April	1.1558 ...	$1.1558 \dots \times 200 = 231.16$
May	1.3767 ...	$1.3767 \dots \times 200 = 275.35$
June	1.2747 ...	$1.2747 \dots \times 200 = 254.96$
July	1.1274 ...	$1.1274 \dots \times 200 = 225.50$
August	1.0594 ...	$1.0594 \dots \times 200 = 221.90$
September	0.9384 ...	$0.9384 \dots \times 200 = 186.97$
October	0.8271 ...	$0.8271 \dots \times 200 = 165.44$
November	0.8045 ...	$0.8045 \dots \times 200 = 160.91$
December	0.9518 ...	$0.9518 \dots \times 200 = 190.37$

NOTES

1. The average of all of the data is found by adding up all of the observations in the time series and dividing by the total number of observations. In this example,

$$\text{Average of all Data} = \frac{2118}{24} = 88.25.$$

2. To calculate the seasonal index for each month, average the observations for the month and divide by the average of all of the data. For example, the average for January is

$$\text{Average for January} = \frac{70 + 61}{2} = 65.5. \text{ Then the seasonal index for}$$

$$\text{January is } \text{Seasonal Index for January} = \frac{65.5}{88.25} = 0.7422 \dots$$

3. If a season has a seasonal index greater than 1, the season is above average. If a season has a seasonal index less than 1, the season is below average. Because April, May, June, July, and August have seasonal indices greater than 1, they are above-average seasons. Similarly, because January, February, March, September, October, November, and December have seasonal indices less than 1, they are below-average seasons.

4. When calculating the forecast for each season, we need to know the “average”, unadjusted forecast for a season, not the total for the entire year. In this example, the total forecasted sales for year 3 is 2,400. Because there are 12 seasons (months) per year, the “average” per season is $\frac{2,400}{12} = 200$. This 200 “average” per season is then adjusted by multiplying by the corresponding seasonal index for each season to find the forecast. For example, the forecast for July is $200 \times 1.1274 \dots = 225.50$.

NOTE

The sum of the seasonal indices equals the number of seasons. In the above example, the sum of the seasonal indices equals 12, the number of seasons.

TRY IT

The table below shows the quarterly sales figures (in \$1,000,000s) for a local business.

Quarter	Year 1	Year 2	Year 3
1	108	114	105
2	125	116	135
3	161	148	150
4	154	163	165

1. Calculate the seasonal index for each quarter.
2. Identify the average, below-average, and above-average seasons.
3. Suppose the business predicts sales of \$600,000,000 for year 4. Calculate the forecast for each quarter of year 4.

Click to see Solution

1. The average of all of the data is 137.

Quarter	Year 1	Year 2	Year 3	Average per Quarter	Seasonal Index
1	108	114	105	109	0.7956...
2	125	116	135	125.333...	0.9148...
3	161	148	150	153	1.1167...
4	154	163	165	160.666...	1.1727...

2. Quarters 1 and 2 are below-average seasons. Quarters 3 and 4 are above-average seasons.

3.

Quarter	Forecast for Year 4
1	$0.7956 \dots \times 150,000,000 = \$119,343,065.69$
2	$0.9148 \dots \times 150,000,000 = \$137,226,277.37$
3	$1.1167 \dots \times 150,000,000 = \$167,518,248.18$
4	$1.1727 \dots \times 150,000,000 = \$175,912,408.76$

Exercises

1. John runs a small business selling ski and snowboard equipment. The quarterly sales (in \$1000s) for the past four years are recorded in the table below.

Quarter	Year 1	Year 2	Year 3	Year 4
1	183	208	189	213
2	220	207	191	206
3	139	117	128	119
4	100	132	128	133

- Calculate the seasonal index for each quarter.
- Suppose John forecasts sales totalling \$900,000 in year 5. What is the forecast for each quarter of year 5?
- Identify the average, above-average, and below-average seasons.

Click to see Answer

Quarter	Seasonal Index	Forecast for Year 4	Average, Above, or Below
1	1.2139...	\$273,134.33	Above
2	1.2613...	\$283,811.71	Above
3	0.7699...	\$173,249.14	Below
4	0.7546...	\$169,804.82	Below

2. A local garden supply store recorded the number of bags of fertilizer (in 1000s) that it sold each month for three years. The data is recorded in the table below.

Month	Year 1	Year 2	Year 3
January	1	3	2
February	2	3	4
March	3	2	3
April	15	11	13
May	13	11	12
June	10	12	9
July	11	9	10
August	9	8	8
September	9	8	12
October	4	4	4
November	4	4	5
December	2	2	1

- Calculate the seasonal index for each month.
- Suppose the store predicts it will sell 114,000 bags of fertilizer in year 4. What is the forecast for each month of year 4?
- Identify the average, above-average, and below-average seasons.

Click to see Answer

Month	Seasonal Index	Forecast for Year 4	Average, Above, or Below
January	0.2962 ...	2,814.82	Below
February	0.4444 ...	4,222.22	Below
March	0.3950 ...	3,753.09	Below
April	1.9259 ...	18,296.30	Above
May	1.7777 ...	16,888.89	Above
June	1.5308 ...	14,543.21	Above
July	1.4814 ...	14,074.07	Above
August	1.2345 ...	11,728.40	Above
September	1.4320 ...	13,604.94	Above
October	0.5925 ...	5,629.63	Below
November	0.6419 ...	6,098.77	Below
December	0.2469 ...	2,345.68	Below

3. A local coffee shop recorded the number of coffees (in 100s) that they sell each day over a six-week period. The data is recorded in the table below.

Day	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
Sunday	9	9	9	11	7	9
Monday	6	5	4	3	5	7
Tuesday	5	5	6	5	7	3
Wednesday	7	6	7	4	7	8
Thursday	7	7	6	3	5	3
Friday	6	4	7	7	3	4
Saturday	8	10	10	10	10	8

- Calculate the seasonal index for each day.
- Suppose the store predicts it will sell 5250 coffees in week 7. What is the forecast for each day of year 7?
- Identify the average, above-average, and below-average seasons.

Click to see Answer

Day	Seasonal Index	Forecast for Week 7	Average, Above, or Below
Sunday	1.3897...	1,042.28	Above
Monday	0.7720...	579.04	Below
Tuesday	0.7977...	598.35	Below
Wednesday	1.003...	752.76	Above
Thursday	0.7977...	598.35	Below
Friday	0.7977...	598.35	Below
Saturday	1.4417...	1,080.88	Above

14.5 REGRESSION MODELS

LEARNING OBJECTIVES

- Construct forecast models using simple linear regression and multiple regression.

A **trend pattern** is the long-term movement or general direction of the data over a period of time. When a trend is present in a time series, the smoothing models and seasonal indices discussed earlier in this chapter are not able to capture the trend component in the time series, which results in an inaccurate forecast. Instead, we need to use a forecast model that accounts for the trend effect. Although trends do not have to follow a linear model, we will focus on two models that capture the linear trend in a time series.

Linear Trend Projection

Previously, we learned about simple linear regression, which models the linear relationship between an independent variable x and a dependent variable y . The equation for the regression line is

$$\hat{y} = b_0 + b_1x$$

\hat{y} = estimated value of y

x = value of the independent variable

b_0 = y -intercept of the line

b_1 = slope of the line

We can apply simple linear regression to a time series to model the linear trend in the time series, creating a **linear trend projection**. A **trend line** is the simple linear regression equation in which

the independent variable x is the time period and \hat{y} is the forecasted value. To emphasize that the independent variable represents time in a trend line model, we will replace the variable x with t .

The equation for the trend line is

$$\hat{y} = b_0 + b_1t$$

\hat{y} = forecasted value of y

t = time period

b_0 = y -intercept of the line

b_1 = slope of the line

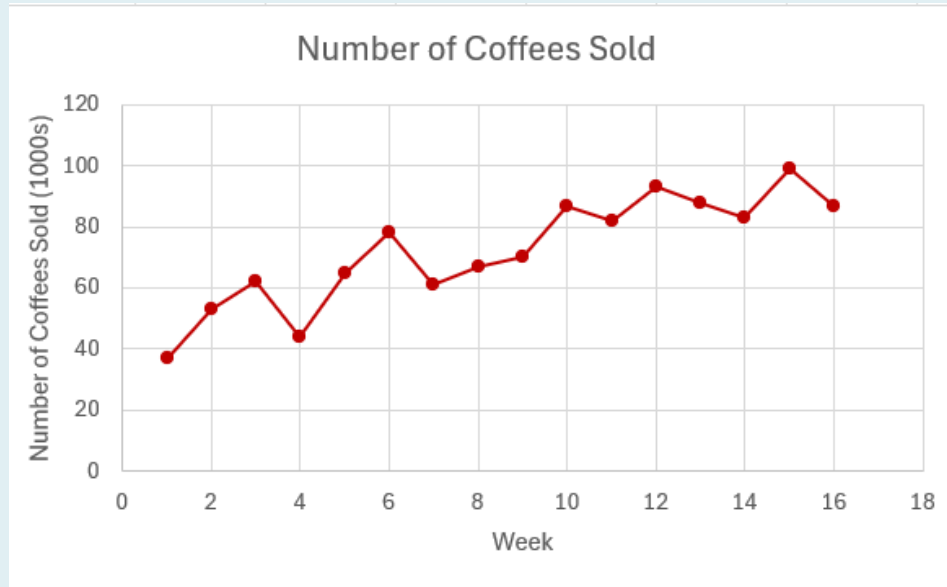
The values for the slope b_1 and the y -intercept b_0 are calculated the same as any simple linear regression equation, using the built-in **slope** and **intercept** functions in Excel.

EXAMPLE

Jane just opened a small coffee shop, located on her town's main street. Jane recorded the number of coffees sold (in 1000s) each week for the first 16 weeks since opening the store.

Week	Number of Coffees Sold (1000s)
1	37
2	53
3	62
4	44
5	65
6	78
7	61
8	67
9	70
10	87
11	82
12	93
13	88
14	83
15	99
16	87

The time series plot, shown below, shows an upward linear trend. So, the time series has a trend component.



1. Create a linear trend projection for this time series.
2. Interpret the slope of the linear trend projection.
3. Use the linear trend projection to forecast the number of coffees sold for week 17.
4. Calculate the *MAD* for this forecast.

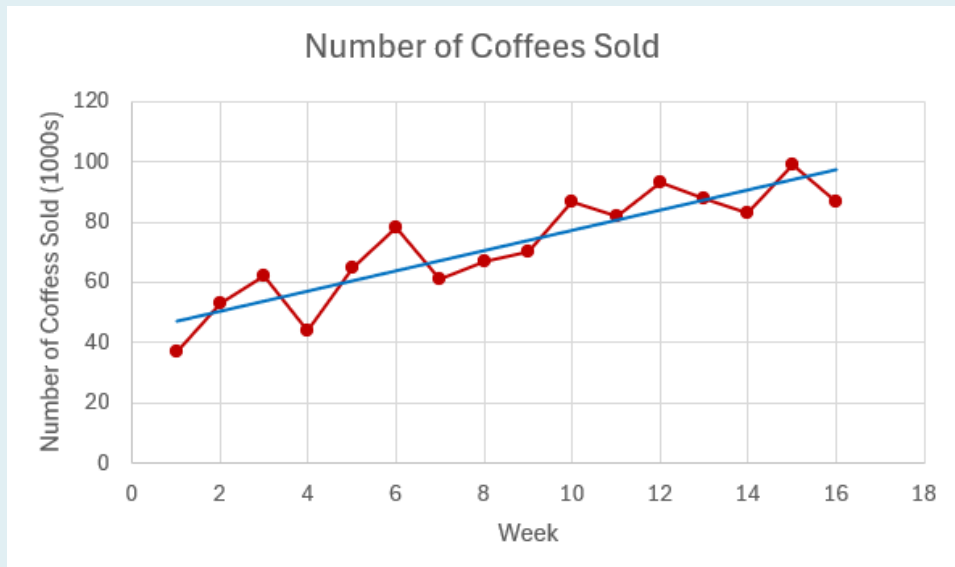
Solution

1. Using the built-in functions in Excel to calculate the slope and intercept of the trend line, the linear trend projection is:

$$\hat{y} = 43.85 + 3.341176 \dots t$$

\hat{y} = forecasted number of coffees in 1000s

t = time period



2. With each additional week, the number of coffees sold increases by 3,341.18.
3. Substitute $t = 17$ into the trend line equation:

$$\begin{aligned}\hat{y} &= 43.85 + 3.341176 \dots \times 17 \\ &= 100.65\end{aligned}$$

The forecasted number of coffees sold in week 17 is 100,650.

4.

Week	Number of Coffees Sold (1000s)	Forecasted Value	Error
1	37	47.191...	10.191...
2	53	50.532...	2.467...
3	62	53.873...	8.126...
4	44	57.214...	13.214...
5	65	60.555...	4.444...
6	78	63.897...	14.102...
7	61	67.238...	6.238...
8	67	70.579...	3.579...
9	70	73.920...	3.920...
10	87	77.261...	9.738...
11	82	80.602...	1.397...
12	93	83.944...	9.055...
13	88	87.285...	0.714...
14	83	90.626...	7.626...
15	99	93.967...	5.032...
16	87	97.308...	10.308...
		Sum	110.158...

$$\begin{aligned}
 MAD &= \frac{\sum |\text{forecast error}|}{\text{number of forecast errors}} \\
 &= \frac{110.158...}{16} \\
 &= 6.885
 \end{aligned}$$

NOTES

1. The slope has the same units as the data in the time series. In this case, the data is the number of coffees in 1000s. So the slope of $3.341176...$ is in 1000s, and is actually $3.341176... \times 1000 = 3,341.17...$

2. The number of coffees sold is given in 1000s, so the forecasted values \hat{y} are also in 1000s. In the forecasted value for week 17, $\hat{y} = 100.65$. This number is in 1000s, so the forecasted number of coffees for week 17 is $100.65 \times 1000 = 100,650$.
3. In the calculation of the *MAD*, first, find the forecasted value for each week in the time series by substituting the time period into the trend line equation and calculating out \hat{y} . For example, the forecasted value for week 3 is

$$\hat{y} = 43.85 + 3.341176 \dots \times 3 = 53.873 \dots$$

NOTES

1. Because the trend line is just an application of simple linear regression, we can apply the same measures used in simple linear regression analysis, such as correlation, coefficient of determination, and standard error of the estimate, to assess how well the trend line fits the time series data.
2. Unlike the smoothing models, which can only forecast the next time period in the time series, trend lines can be used to forecast further into the future. In the above example, we could use the trend line to create a forecast for week 20 or week 30. However, the further away from the time series data, the less reliable the forecast.

TRY IT

The yearly operating budget (in \$1,000,000s) for a local city is provided in the table below.

Year	Budget (\$1,000,000s)
1	3.14
2	3.27
3	3.33
4	3.45
5	3.68
6	3.76
7	3.98
8	4.05
9	4.12
10	4.47
11	4.65
12	4.78
13	4.82
14	4.85
15	4.93
16	5.02

1. Create a linear trend projection for this time series.
2. Interpret the slope of the linear trend projection.
3. Use the linear trend projection to forecast the number of coffees sold for year 20.

Click to see Solution

$$\hat{y} = 2.987 \dots + 0.136 \dots t$$

1. \hat{y} = forecasted budget in \$1,000,000s
 t = time period

2. For each additional year, the budget increases by \$136,058.82.
3. The forecast for year 20 is \$5,708,426.47.

Multiple Regression

The linear trend projection discussed above can only handle the trend component of a time series. But many time series include both trend and seasonal components. In such cases, we need to use a forecasting model that incorporates both of these components. One possible model that addresses both trend and seasonal components is a multiple regression model.

Previously, we learned about multiple regression, which models the linear relationship between one dependent variable and several independent variables. The equation for the multiple regression model is:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

where \hat{y} is the predicted value of y , x_1, x_2, \dots, x_k are the independent variables, b_0, b_1, \dots, b_k are the regression coefficients, and k is the number of independent variables.

As with the linear trend projection, \hat{y} in a multiple regression model for a time series is the forecasted value. The number of independent variables in a multiple regression model for a time series equals the number of seasons. For example, a regression model for a time series using quarterly data will have four independent variables because the time series has four seasons. Similarly, a regression model for a time series using monthly data will have twelve independent variables because the time series has twelve seasons. The independent variable x_1 is the time period, similar to t in the linear trend projection model above. The remaining independent variables are **dummy variables** to indicate the season.

NOTE

A **dummy variable** is a variable that is assigned a value of 1 if a particular condition is met and

a value of 0 otherwise. In the context of a multiple regression model for a time series, dummy variables are used for the different seasons.

For example, suppose a time series uses quarterly data. The multiple regression model has four independent variables: x_1, x_2, x_3, x_4 . The variable x_1 is for the time period. The variable x_2 is a dummy variable for quarter 2 and equals 1 when forecasting a quarter 2 value and 0 otherwise. The variable x_3 is a dummy variable for quarter 3 and equals 1 when forecasting a quarter 3 value and 0 otherwise. The variable x_4 is a dummy variable for quarter 4 and equals 1 when forecasting a quarter 4 value and 0 otherwise. When $x_2 = x_3 = x_4 = 0$, then the quarter being forecasted is a quarter 1 value.

For a time series with quarterly data, the multiple regression model is

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$$

\hat{y} = forecasted value

x_1 = time period

x_2 = 1 if quarter 2 or 0 otherwise

x_3 = 1 if quarter 3 or 0 otherwise

x_4 = 1 if quarter 4 or 0 otherwise

For a time series with monthly data, the multiple regression model is

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_{12}x_{12}$$

\hat{y} = forecasted value

x_1 = time period

x_2 = 1 if month 2 or 0 otherwise

x_3 = 1 if month 3 or 0 otherwise

\vdots

x_{12} = 1 if month 12 or 0 otherwise

In general, for a time series with k seasons, there are k independent variables. The variable x_1 is the time period variable, and the variables x_2, x_3, \dots, x_k are dummy variables.

NOTE

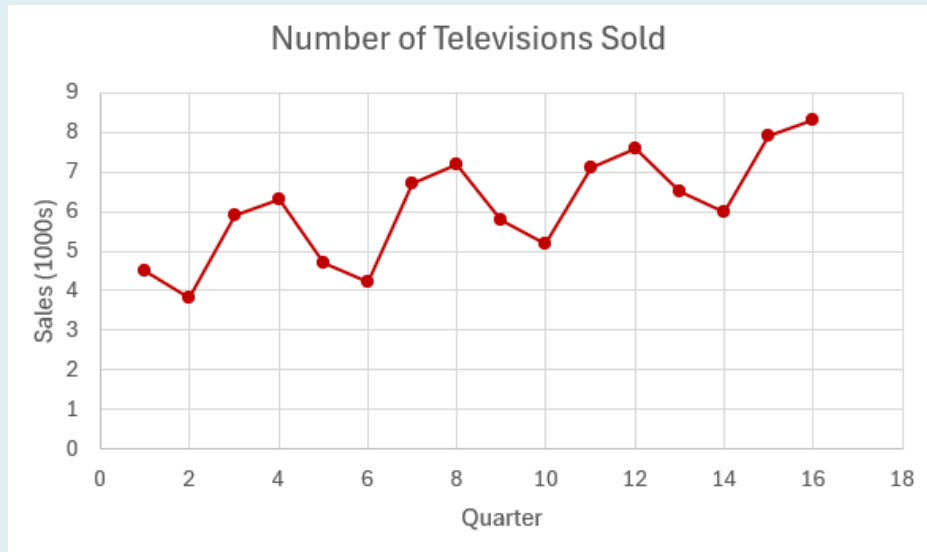
The independent variable that corresponds to the time period is not a dummy variable. The choice of which independent variable corresponds to the time period is arbitrary. In a multiple regression model, the number of independent variables equals the number of seasons. Any of those independent variables can be selected as the time period variable and then the remaining independent variables become dummy variables. The coefficients in the model will be different, but the forecast will be the same regardless of which independent variable is used for the time period. For convenience and consistency, we will designate x_1 as the time period variable.

EXAMPLE

A company makes television sets. The number of sets sold (in 1000s) each quarter for the past four years is recorded in the table below.

Year	Quarter	Sales (1000s)
1	1	4.5
	2	3.8
	3	5.9
	4	6.3
2	5	4.7
	6	4.2
	7	6.7
	8	7.2
3	9	5.8
	10	5.2
	11	7.1
	12	7.6
4	13	6.5
	14	6
	15	7.9
	16	8.3

The time series plot, shown below, indicates that there are drops in sales in the second quarter of each year and that there are peaks in sales in the third and fourth quarters of each year. The time series plot also shows an upward linear trend. So, the time series has both seasonal and trend components.



1. Create a multiple regression model for this time series.
2. Use the multiple regression model to forecast the number of televisions sold for each quarter for year 5.
3. Calculate the *MAPE* for this forecast.

Solution

1. The time series has four seasons/quarters, so the multiple regression model will have four independent variables: x_1 for the time period and dummy variables x_2 , x_3 , x_4 for quarters 2, 3 and 4. To use Excel's regression analysis, we need to set up a table, shown below, containing the values of the independent variables for the data. The values for the time period variable x_1 are the time period values. We need to add three additional columns for the dummy variables x_2 , x_3 , x_4 . The values in these additional columns are 1 for the corresponding quarter or 0 otherwise. For example, in the Quarter 2 column, there is a 1 in each row that corresponds to a Quarter 2 (quarters 2, 6, 10, and 14) in the data and a 0 everywhere else. In the Quarter 3 column, there is a 1 in each row that corresponds to a Quarter 3 (quarters 3, 7, 11, and 15) in the data and a 0 everywhere else. In the Quarter 4 column, there is a 1 in each row that corresponds to a Quarter 4 (quarters 4, 8, 12, and 16) in the data and a 0 everywhere else.

Time Period	Quarter 2	Quarter 3	Quarter 4	Sales (1000s)
1	0	0	0	4.5
2	1	0	0	3.8
3	0	1	0	5.9
4	0	0	1	6.3
5	0	0	0	4.7
6	1	0	0	4.2
7	0	1	0	6.7
8	0	0	1	7.2
9	0	0	0	5.8
10	1	0	0	5.2
11	0	1	0	7.1
12	0	0	1	7.6
13	0	0	0	6.5
14	1	0	0	6
15	0	1	0	7.9
16	0	0	1	8.3

Using this table, apply Excel's regression analysis to find the multiple regression model:

$$\hat{y} = 4.171875 + 0.17875x_1 - 0.746875x_2 + 1.18125x_3 + 1.459375x_4$$

\hat{y} = forecasted sales in 1000s

x_1 = time period

x_2 = 1 if quarter 2 or 0 otherwise

x_3 = 1 if quarter 3 or 0 otherwise

x_4 = 1 if quarter 4 or 0 otherwise

2. Forecast for quarter 1 of year 5 (time period 17):

$$\begin{aligned}\hat{y} &= 4.171875 + 0.17875 \times 17 - 0.746875 \times 0 + 1.18125 \times 0 + 1.459375 \times 0 \\ &= 7.09375\end{aligned}$$

The forecasted sales for quarter 1 of year 5 is 7,093.75.

Forecast for quarter 2 of year 5 (time period 18):

$$\begin{aligned}\hat{y} &= 4.171875 + 0.17875 \times 18 - 0.746875 \times 1 + 1.18125 \times 0 + 1.459375 \times 0 \\ &= 6.51875\end{aligned}$$

The forecasted sales for quarter 2 of year 5 is 6,518.75.

Forecast for quarter 3 of year 5 (time period 19):

$$\begin{aligned}\hat{y} &= 4.171875 + 0.17875 \times 19 - 0.746875 \times 0 + 1.18125 \times 1 + 1.459375 \times 0 \\ &= 8.61875\end{aligned}$$

The forecasted sales for quarter 3 of year 5 is 8,618.75.

Forecast for quarter 4 of year 5 (time period 20):

$$\begin{aligned}\hat{y} &= 4.171875 + 0.17875 \times 20 - 0.746875 \times 0 + 1.18125 \times 0 + 1.459375 \times 1 \\ &= 9.06875\end{aligned}$$

The forecasted sales for quarter 4 of year 5 is 9,068.75.

3.

Quarter	Sales (1000s)	Forecasted Value	Percent Error
1	4.5	4.34375	0.0347...
2	3.8	3.76875	0.0082...
3	5.9	5.86875	0.0052...
4	6.3	6.31875	0.0029...
5	4.7	5.03125	0.0704...
6	4.2	4.45625	0.0610...
7	6.7	6.55625	0.0214...
8	7.2	7.00625	0.0269...
9	5.8	5.71875	0.0140...
10	5.2	5.14375	0.0108...
11	7.1	7.24375	0.0202...
12	7.6	7.69375	0.0123...
13	6.5	6.40625	0.0144...
14	6	5.83125	0.0281...
15	7.9	7.93125	0.0039...
16	8.3	8.38125	0.0097...
		Sum	0.3447...

$$\begin{array}{l} \text{MAPE} = \frac{\sum |\text{percent error}|}{\text{number of}} \\ \text{forecast errors}} \times 100\% = \frac{0.3447}{16} \times 100\% = 2.15\% \end{array}$$

NOTES

1. The number of televisions sold is given in **1000s**, so the forecasted values \hat{y} are also in **1000s**. In the forecasted value for quarter 1 of year 5, $\hat{y} = 7.09375$. This number is in **1000s**, so the forecasted number of televisions sold for quarter 1 of year 5 is $7.09375 \times 1000 = 7,093.75$.
2. In the calculation of the *MAPE*, first find the forecasted value for each quarter in the time series by substituting the time period and corresponding values of the dummy variables into the multiple regression model and calculating out \hat{y} . For example, the forecasted value for quarter 7 is

$$\hat{y} = 4.171875 + 0.17875 \times 7 - 0.746875 \times 0 + 1.18125 \times 1 + 1.459375 \times 0 = 6.55625$$
.

TRY IT

The number of daily calls to emergency services in a small town is given in the table below.

Week	Day	Number of Calls
1	Sunday	38
	Monday	25
	Tuesday	26
	Wednesday	24
	Thursday	24
	Friday	29
	Saturday	31
2	Sunday	34
	Monday	25
	Tuesday	23
	Wednesday	26
	Thursday	24
	Friday	27
	Saturday	29
3	Sunday	30
	Monday	24
	Tuesday	22
	Wednesday	23
	Thursday	21
	Friday	26
	Saturday	27

1. Create a multiple regression model for this time series.
2. Use the multiple regression model to forecast the number of calls for each day of week 4.

Click to see Solution

$$\hat{y} = 35.959 \dots - 0.244 \dots x_1 - 9.088 \dots x_2 - 9.843 \dots x_3 - 8.931 \dots x_4 \\ - 10.020 \dots x_5 - 5.442 \dots x_6 - 3.530 \dots x_7$$

\hat{y} = forecasted number of calls

x_1 = time period

1. x_2 = 1 if Monday or 0 otherwise
 x_3 = 1 if Tuesday or 0 otherwise
 x_4 = 1 if Wednesday or 0 otherwise
 x_5 = 1 if Thursday or 0 otherwise
 x_6 = 1 if Friday or 0 otherwise
 x_7 = 1 if Saturday or 0 otherwise

2. Forecast for Sunday of week 4 (time period 22):

$$\hat{y} = 35.959 \dots - 0.244 \dots \times 22 - 9.088 \dots \times 0 - 9.843 \dots \times 0 - 8.931 \dots \times 0 \\ - 10.020 \dots \times 0 - 5.442 \dots \times 0 - 3.530 \dots \times 0 \\ = 30.57$$

Forecast for Monday of week 4 (time period 23):

$$\hat{y} = 35.959 \dots - 0.244 \dots \times 23 - 9.088 \dots \times 1 - 9.843 \dots \times 0 - 8.931 \dots \times 0 \\ - 10.020 \dots \times 0 - 5.442 \dots \times 0 - 3.530 \dots \times 0 \\ = 21.24$$

Forecast for Tuesday of week 4 (time period 24):

$$\hat{y} = 35.959 \dots - 0.244 \dots \times 24 - 9.088 \dots \times 0 - 9.843 \dots \times 1 - 8.931 \dots \times 0 \\ - 10.020 \dots \times 0 - 5.442 \dots \times 0 - 3.530 \dots \times 0 \\ = 20.24$$

Forecast for Wednesday of week 4 (time period 25):

$$\hat{y} = 35.959 \dots - 0.244 \dots \times 25 - 9.088 \dots \times 0 - 9.843 \dots \times 0 - 8.931 \dots \times 1 \\ - 10.020 \dots \times 0 - 5.442 \dots \times 0 - 3.530 \dots \times 0 \\ = 20.9$$

Forecast for Thursday of week 4 (time period 26):

$$\hat{y} = 35.959 \dots - 0.244 \dots \times 26 - 9.088 \dots \times 0 - 9.843 \dots \times 0 - 8.931 \dots \times 0 \\ - 10.020 \dots \times 1 - 5.442 \dots \times 0 - 3.530 \dots \times 0 \\ = 19.57$$

Forecast for Friday of week 4 (time period 27):

$$\begin{aligned}\hat{y} &= 35.959 \dots - 0.244 \dots \times 27 - 9.088 \dots \times 0 - 9.843 \dots \times 0 - 8.931 \dots \times 0 \\ &\quad - 10.020 \dots \times 0 - 5.442 \dots \times 1 - 3.530 \dots \times 0 \\ &= 23.9\end{aligned}$$

Forecast for Saturday of week 4 (time period 28):

$$\begin{aligned}\hat{y} &= 35.959 \dots - 0.244 \dots \times 28 - 9.088 \dots \times 0 - 9.843 \dots \times 0 - 8.931 \dots \times 0 \\ &\quad - 10.020 \dots \times 0 - 5.442 \dots \times 0 - 3.530 \dots \times 1 \\ &= 25.57\end{aligned}$$

Exercises

1. A company that makes kitchen appliances has just launched a new convection oven on the market. Using the weekly sales (in thousands of units) for the eight weeks the oven has been on the market, the company created the linear trend model $\hat{y} = 19 + 1.06t$ where t is the time period and \hat{y} is the forecasted sales (in thousands).
 - a. Interpret the slope of the trend line equation.
 - b. Use the trend line to forecast the sales for week 9.
 - c. Use the trend line to forecast the sales for week 12.

Click to see Answer

- a. For each additional week, the sales increase by 1,060 units.
 - b. 28,540
 - c. 31,720
2. The number of cars sold each year for a local used car dealership is recorded in the table below.

Year	Number of Cars Sold
1	405
2	370
3	289
4	365
5	302
6	315
7	320
8	265
9	278
10	211

- Create a linear trend projection forecast model for this time series.
- Interpret the slope of the trend line equation.
- Calculate the correlation coefficient for the trend line equation.
- Interpret the correlation coefficient found in part c.
- Use the trend line to forecast the sales for year 11.
- Calculate the *MAPE* for this forecast.

Click to see Answer

- $\hat{y} = 399.7333 - 15.9515t$ where t is the year and \hat{y} is the forecasted number of cars sold.
- For each additional year, the number of cars sold decreases by 15.95.
- 0.8495
- There is a strong, negative linear relationship between year and sales.
- 224.27
- 7.93%

- An industrial cleaning supply company recorded the monthly sales (in \$1000s) of their industrial vacuum cleaner.

Month	Sales (\$1000s)
1	12
2	14
3	11
4	15
5	16
6	14
7	17
8	19
9	15
10	17
11	18
12	18
13	21

- Create a linear trend projection forecast model for this time series.
- Interpret the slope of the trend line equation.
- Calculate the correlation coefficient for the trend line equation.
- Interpret the correlation coefficient found in part c.
- Use the trend line to forecast the sales for month 15.
- Calculate the MSE for this forecast.

Click to see Answer

- $\hat{y} = 11.653 + 0.609t$ where t is the month and \hat{y} is the forecasted sales.
- For each additional month, the sales increase by \$609.89.
- 0.8445
- There is a strong, positive linear relationship between month and sales.
- \$20,802.20
- 2.094

- The number of surgeries performed each quarter at a small hospital is recorded in the table below.

Quarter	Number of Surgeries
1	65
2	57
3	52
4	49
5	51
6	50
7	43
8	42
9	44
10	40

- Create a linear trend projection forecast model for this time series.
- Interpret the slope of the trend line equation.
- Use the trend line to forecast the sales for quarter 13.
- Calculate the MAD for this forecast.

Click to see Answer

- $\hat{y} = 62.133 - 2.333t$ where t is the quarter and \hat{y} is the forecasted number of surgeries.
 - For each additional quarter, the number of surgeries decreases by 2.333.
 - 31.8
 - 2.333
5. Using five years of quarterly revenue data (in \$1, 000, 000s), a national restaurant chain created the following multiple regression model to forecast the quarterly revenue.

$$\hat{y} = 26.71 + 3.34x_1 - 4.54x_2 - 5.89x_3 - 2.23x_4$$

\hat{y} = quarterly revenue in \$1, 000, 000s

x_1 = time period

x_2 = 1 if quarter 2 or 0 otherwise

x_3 = 1 if quarter 3 or 0 otherwise

x_4 = 1 if quarter 4 or 0 otherwise

- Use the multiple regression model to forecast the quarterly revenue for quarter 1 of year

- 6.
- b. Use the multiple regression model to forecast the quarterly revenue for quarter 3 of year 6.
- c. Use the multiple regression model to forecast the quarterly revenue for quarter 2 of year 7.
- d. Use the multiple regression model to forecast the quarterly revenue for quarter 4 of year 7.

Click to see Answer

- a. \$96,850,000
- b. \$97,640,000
- c. \$109,010,000
- d. \$118,000,000

6. A major source of revenue for a country is the sales tax on goods and services. The quarterly revenue, in millions of dollars, generated by the sales tax is recorded in the table below.

Quarter	Revenue from Sales Tax (\$1,000,000s)
1	212
2	239
3	299
4	325
5	248
6	277
7	320
8	340
9	270
10	290
11	340
12	389

- a. Create a multiple regression forecast model for this time series.
- b. Use the multiple regression model to forecast the revenue for quarter 13.
- c. Use the multiple regression model to forecast the revenue for quarter 15.

- d. Use the multiple regression model to forecast the revenue for quarter 20.
- e. Calculate the MAD for this forecast.

Click to see Answer

$$\hat{y} = 225.476 + 8.485x_1 - 8.722x_2 + 18.091x_3 + 47.105x_4$$

\hat{y} = forecasted revenue

- a. x_1 = time period
 x_2 = 1 if quarter 2 or 0 otherwise
 x_3 = 1 if quarter 3 or 0 otherwise
 x_4 = 1 if quarter 4 or 0 otherwise
- b. \$336,786,699.11
- c. \$371,848,134.63
- d. \$443,289,740.47
- e. 20.06

7. A customer comment line is staffed from 8:00 am to 4:30 pm, Monday to Friday. The number of calls received every day for the past five weeks are recorded in the table below.

Week	Day	Number of Calls
1	Monday	35
	Tuesday	17
	Wednesday	18
	Thursday	20
	Friday	29
2	Monday	33
	Tuesday	16
	Wednesday	17
	Thursday	16
	Friday	28
3	Monday	29
	Tuesday	15
	Wednesday	13
	Thursday	14
	Friday	23
4	Monday	27
	Tuesday	15
	Wednesday	13
	Thursday	15
	Friday	25
5	Monday	27
	Tuesday	14
	Wednesday	13
	Thursday	12
	Friday	22

- Create a multiple regression forecast model for this time series.
- Use the multiple regression model to forecast the number of calls for each day in week 6.
- Calculate the $MAPE$ for this forecast.

Click to see Answer

$$\hat{y} = 33.588 - 0.308x_1 - 14.492x_2 - 14.784x_3 - 13.8765x_4 - 3.568x_5$$

\hat{y} = forecasted number of calls

a. x_1 = time period

x_2 = 1 if Tuesday or 0 otherwise

x_3 = 1 if Wednesday or 0 otherwise

x_4 = 1 if Thursday or 0 otherwise

x_5 = 1 if Friday or 0 otherwise

b. Monday = 25.58, Tuesday = 10.78, Wednesday = 10.18, Thursday = 10.78,
Friday = 20.78

c. 5.69\%

8. Joe runs a lawn and garden supply store. The monthly sales (in \$100,000) for the past two years are recorded in the table below.

Year	Month	Sales (\$100, 000s)
1	January	3
	February	4
	March	9
	April	13
	May	25
	June	23
	July	22
	August	20
	September	10
	October	6
	November	4
	December	3
2	January	2
	February	3
	March	2
	April	7
	May	26
	June	30
	July	32
	August	29
	September	10
	October	6
	November	4
	December	3

- Create a multiple regression forecast model for this time series.
- Use the multiple regression model to forecast the monthly sales for March of year 3.
- Use the multiple regression model to forecast the monthly sales for July of year 3.
- Use the multiple regression model to forecast the monthly sales for September of year 3.
- Use the multiple regression model to forecast the monthly sales for December of year 3.

Click to see Answer

$$\hat{y} = 1.917 + 0.083x_1 + 0.917x_2 + 2.833x_3 + 7.25x_4 + 22.667x_5 + 23.583x_6 + 24x_7 \\ + 21.417x_8 + 6.833x_9 + 2.75x_{10} + 0.667x_{11} - 0.417x_{12}$$

\hat{y} = forecasted sales

x_1 = time period

x_2 = 1 if February or 0 otherwise

x_3 = 1 if March or 0 otherwise

x_4 = 1 if April or 0 otherwise

a.

x_5 = 1 if May or 0 otherwise

x_6 = 1 if June or 0 otherwise

x_7 = 1 if July or 0 otherwise

x_8 = 1 if August or 0 otherwise

x_9 = 1 if September or 0 otherwise

x_{10} = 1 if October or 0 otherwise

x_{11} = 1 if November or 0 otherwise

x_{12} = 1 if December or 0 otherwise

b. \$700,000

c. \$2,850,000

d. \$1,150,000

e. \$450,000

PART XV

STATISTICAL QUALITY CONTROL

As consumers, we take many things into consideration before purchasing a product or service. For example, we often consider the price of the product or service to ensure we are getting good value for our money. But another important factor we consider is **quality**. The quality of a product or service, which is the degree to which a product or service meets its specifications, is a major issue for both the provider of the product or service and the customer. Poor quality may result in the producer recalling the product, which may cost the producer a significant amount of money. Similarly, poor quality may be expensive for the consumer or discourage potential customers from purchasing the product or service altogether.

Not surprisingly, quality is an important part of businesses that manufacture products, ensuring that the product meets the required specifications. Manufacturing firms employ quality control tactics, such as inspections and measurements, to monitor the production process. Among other things, these quality control processes ensure that the product is free of defects, that the product is delivered at the right time and to the right place, and that the product is sold at the right price. When the quality standards are not met, preventive or corrective action is taken to move the production process back into compliance.

But quality is just as important to businesses that provide services, such as banking, healthcare, and education. In the service sector, quality is applied to ensure that the service provided meets the needs and expectations of the customer. In the service industry, quality control tactics, such as monitoring timelines for providing service or customer satisfaction surveys, focus on improving customer service and customer satisfaction.

Although there are other ways to measure quality, this chapter focuses on **statistical process control**. The goal of statistical process control, which utilizes a graphical display called a control chart, is to monitor a process and then determine if the process can continue or if corrective action must be taken to meet the desired level of quality for the process.

CHAPTER OUTLINE

15.1 Control Charts

15.2 Control Charts for Variables

15.3 Control Charts for Attributes

15.1 CONTROL CHARTS

LEARNING OBJECTIVES

- Construct a control chart to monitor the quality of a process.
- Analyze a control chart to determine if a process is in or out of control.

The **quality** of a product or service is the degree to which the product or service meets specifications. Although quality control applies to both products and services, for simplicity, this chapter focuses on products and the manufacturing process used to produce those products. Through a regular process of sampling and inspections of the production process, a decision is made to either continue production or adjust the production process to bring the product up to acceptable quality standards.

Even though a manufacturing process employs the highest quality standards, there are always factors that disrupt the production process and cause a poor quality output. Factors that might result in poor quality include machines or tools breaking down or wearing out, defective or poor quality raw or purchased materials, vibrations altering a machine's settings, or mistakes made by human operators. By monitoring the process and production output, poor quality can be detected early, and corrective action taken to adjust the production process.

All processes contain some degree of variability, but the goal is to keep those variations under control and within acceptable limits. Variations fall into one of two categories—assignable or natural.

- **Assignable variations** are variations in the production process that are not random and controllable. For example, worn-out tools or machines, defects in raw materials, improperly adjusted equipment, and human error are assignable variations. These types of variations can be controlled and corrected. When quality is affected by assignable variations, the process is

considered out-of-control and should be adjusted as soon as possible.

- **Natural or common variations** are variations in the production process that are random and uncontrollable. For example, random variations caused by temperature or humidity are natural variations. These types of variations occur in almost every production process and cannot be controlled by the manufacturer. When quality is affected by natural variations, the process is considered in-control and does not require adjustment.

Statistical process control is a process that sets standards for the product, monitors that the product meets those standards, takes measurements, and corrects quality problems as the product is produced. Of course, it is not possible to assess every single product that is produced. Instead, random samples of the product are taken and examined. If the samples fall within acceptable limits, the process is in-control and is allowed to continue. If the samples fall outside the acceptable limits, the production process is out-of-control, the process is stopped, any assignable variations are identified, and corrective action is taken to resolve those assignable variations.

NOTE

Statistical process control procedures are similar to hypothesis testing. The null hypothesis is that the process is in-control and the alternative hypothesis is that the process is out-of-control. Samples of the process are taken as evidence to either support the claim that the process is in-control (the null hypothesis) or to support the claim that the process is out-of-control (the alternative hypothesis). However, like any hypothesis test, the test may not identify the correct decision. For example, the statistical process control procedures may indicate that the process is out-of-control (the null hypothesis is false) when, in fact, the process is in-control, which is a Type I error in hypothesis testing. Similarly, the statistical process control procedures may indicate that the process is in-control (the null hypothesis is true) when, in fact, the process is out-of-control, which is a Type II error in hypothesis testing.

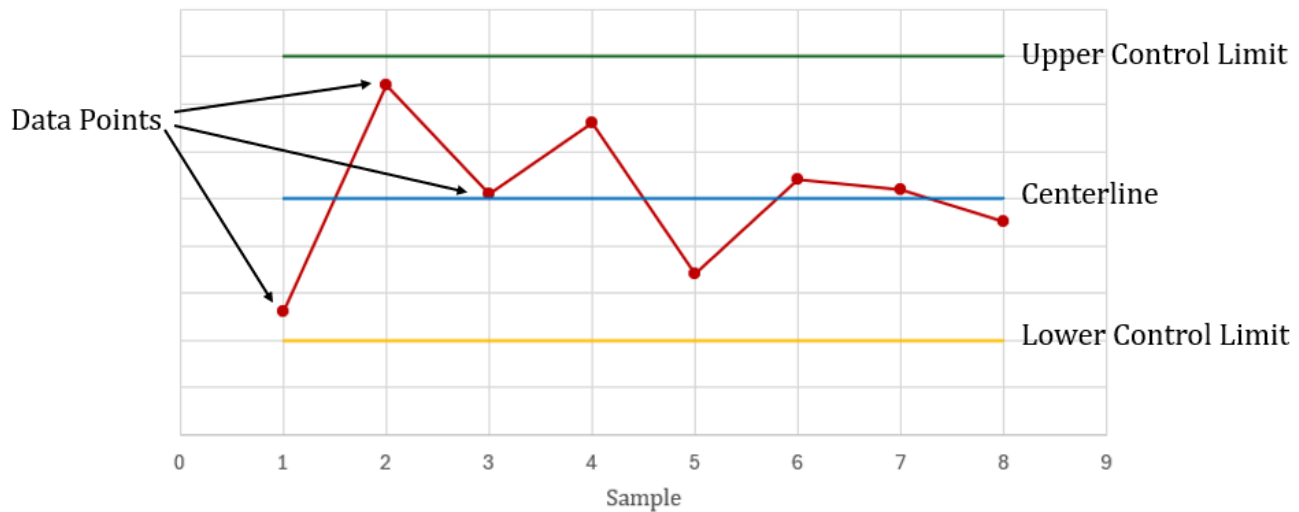
Control Charts

A **control chart** is a statistical tool used to monitor the quality or consistency of a process over time. A control chart is a graphical presentation of data over time. A control chart helps determine whether a process is in a state of statistical control or if there are any (assignable) variations that need corrective actions. For example, suppose the process is monitoring the mean of the samples.

A series of samples are taken at different times, and the mean of each sample is calculated and plotted against time or order of production on the control chart. Typically, small samples are used to monitor the process on a control chart instead of individual items. Generally, individual items contain too much variability, which makes it difficult to identify trends in the process.

A control chart plots data points against time or order of production, similar to a time-series plot. Along with the data points, control charts include three lines—the centerline, the upper control limit, and the lower control limit—that help assess the stability and variability of a process over time.

- **Centerline.** The centerline represents the target or average value of the process being monitored on the control chart. The centerline is a reference point, which allows us to see how the data behaves relative to the expected or desired outcome. For example, suppose a company produces 300 gram bags of ground coffee. As part of the quality control process, the company monitors the average weight of coffee in a sample of bags. The centerline of the control chart is 300 grams, the target weight of each bag of coffee.
- **Upper Control Limit (UCL).** The upper control limit is the highest value allowed for a data point to be considered in-control. The value of the upper control limit is based on a statistical formula that accounts for the variation in the process. Typically, the value of the upper control limit is two or three standard deviations above the centerline in a normally distributed process. The upper control limit helps detect when a process may be out-of-control due to significant variation in the process. Generally, a data point above the upper control limit indicates that the process may be experiencing abnormal variability that requires further investigation. For example, suppose a company produces 300 gram bags of ground coffee. If the upper control limit is set at 305 grams, any sample with a mean weight greater than 305 grams may suggest a problem with the production process.
- **Lower Control Limit (LCL).** The lower control limit is the lowest value allowed for a data point to be considered in-control. The value of the lower control limit is based on a statistical formula that accounts for the variation in the process. Typically, the value of the lower control limit is two or three standard deviations below the centerline in a normally distributed process. The lower control limit helps detect when a process may be out-of-control due to significant variation in the process. Generally, a data point below the lower control limit indicates that the process may be experiencing abnormal variability that requires further investigation. For example, suppose a company produces 300 gram bags of ground coffee. If the lower control limit is set at 295 grams, any sample with a mean weight less than 295 grams may suggest a problem with the production process.



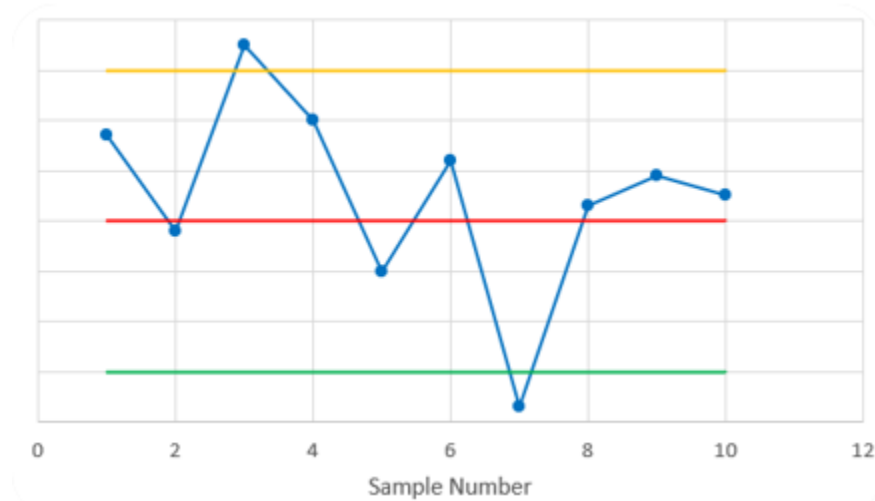
The upper control limit and lower control limit are calculated so that there is a high probability that data points fall in between the two limits when the process is in-control. Together, the centerline, upper control limit, and lower control limit help to determine if the process is in-control or out-of-control where corrective action is required. Data points falling outside the control limits typically signal that the process is out-of-control and that the process needs to be investigated for potential causes.

Interpreting a Control Chart

Interpreting a control chart involves analyzing the data points in relation to the centerline, upper control limit, and lower control limit, as well as recognizing patterns or trends that may indicate a need for corrective action. When a process is in-control, a control chart shows random fluctuations above and below the centerline within the upper and lower control limits. Abnormal or non-random appearing patterns on a control chart may indicate that a process is out-of-control, requiring corrective action.

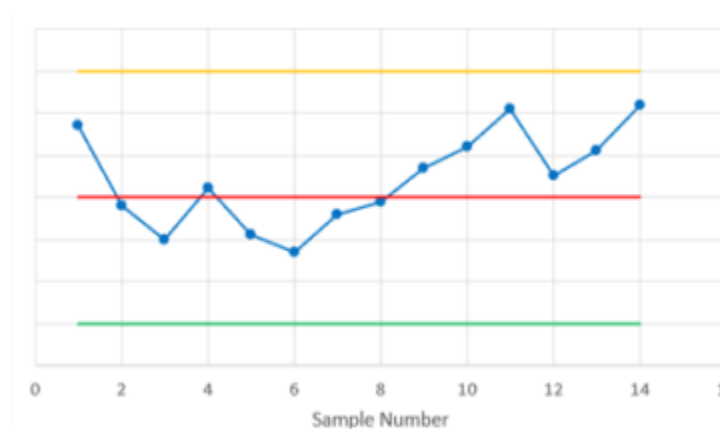
Below is a list of abnormal or unusual control chart patterns that may indicate an out-of-control process.

1. **Data points above the upper control limit or below the lower control limit.** When any data points fall outside the control limits, the process is considered out-of-control. In such cases, some unexpected or abnormal variation may be influencing the process, requiring corrective action.

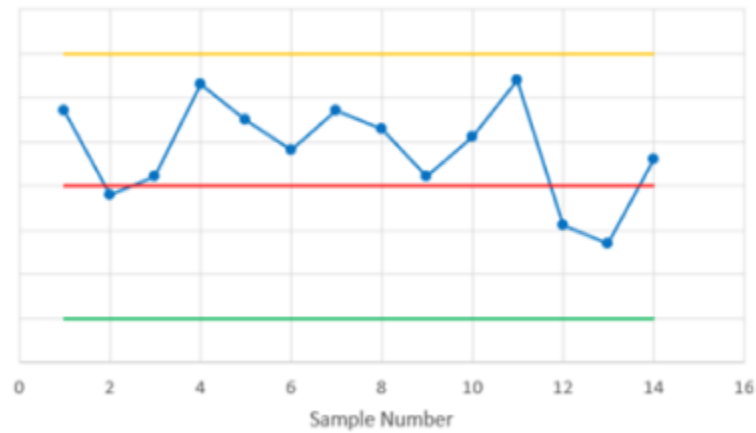


When all of the data points are inside the control limits, the process is typically considered in-control. However, even when all the points are inside the control limits, certain patterns in the data may indicate an out-of-control process.

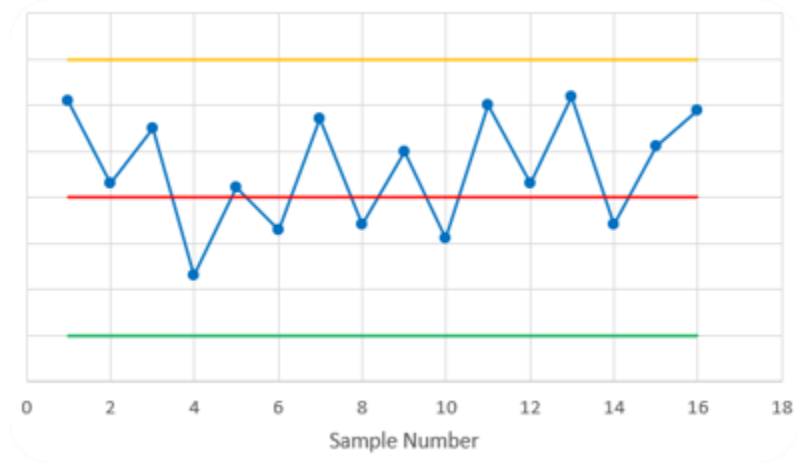
2. **Trends of six or more consecutive increasing or decreasing data points.** Even if the data points are within the control limits, a consistent upward or downward movement of data points over time may indicate that something is changing in the process. In such cases, the process is considered out-of-control, requiring corrective action.



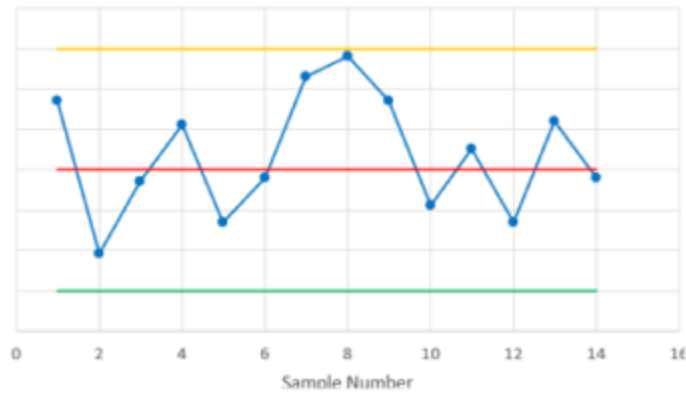
3. **Eight or more consecutive data points are all above or all below the centerline.** A run of a certain number of consecutive data points all on one side of the centerline (either above or below it) may indicate something is changing in the process. In addition to eight or more consecutive data points all above or below the centerline, other patterns that may indicate a run include ten out of eleven data points all above or all below the centerline or twelve out of fourteen data points all above or all below the centerline. In such cases, the process is considered out-of-control, requiring corrective action.



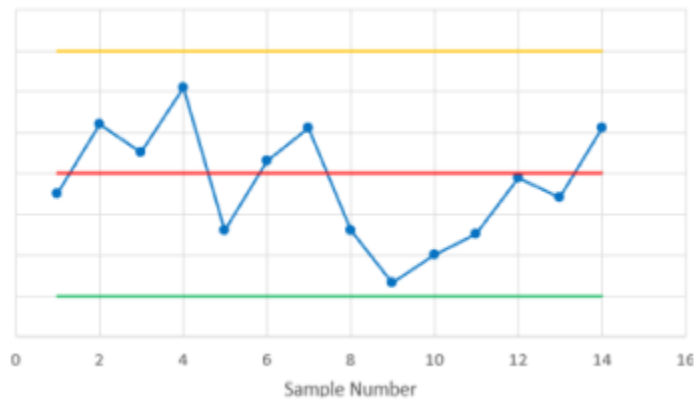
4. **Fourteen or more consecutive points that alternate in direction (increasing then decreasing).** Data points that follow a repeating, alternating pattern of high and low values may indicate something is changing the process. In such cases, the process is considered out-of-control, requiring corrective action.



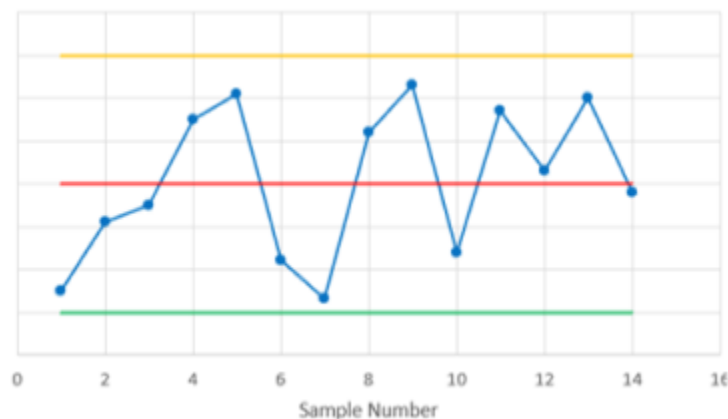
5. **Two out of three consecutive data points are in the outer one-third of the control chart.** When two out of three consecutive data points are more than two standard deviations from the centerline, it signals a potential issue in the process that may need investigation.



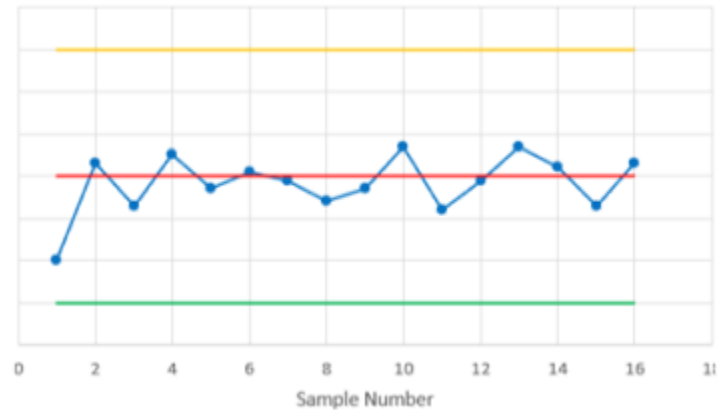
6. **Four out of five consecutive data points are in the outer two-thirds of the control chart.** When four out of five consecutive data points are more than one standard deviation from the centerline, it signals a potential issue in the process that may need investigation.



7. **Eight consecutive data points with none within one standard deviation of the centerline.** When eight consecutive data points are all more than one standard deviation from the centerline and the points are in both directions, it signals a potential issue in the process.

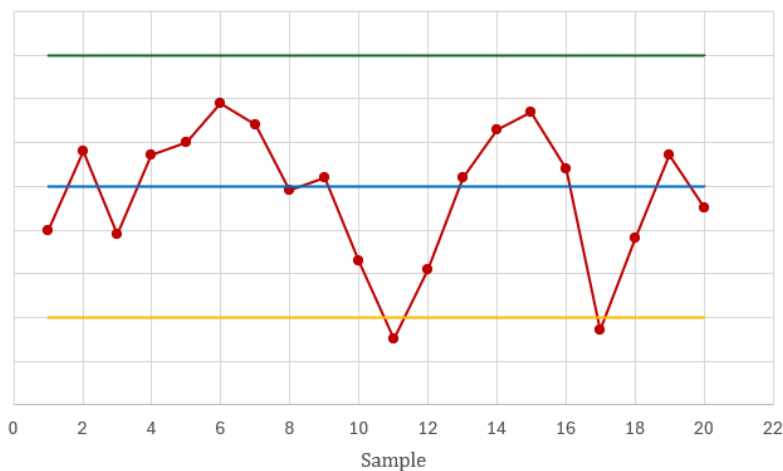


8. **Fifteen consecutive data points are all within one standard deviation of the centerline on either side of the centerline.** When fifteen consecutive data points are all within one standard deviation of the centerline, on either side of the centerline, it signals a potential issue in the process.



Exercises

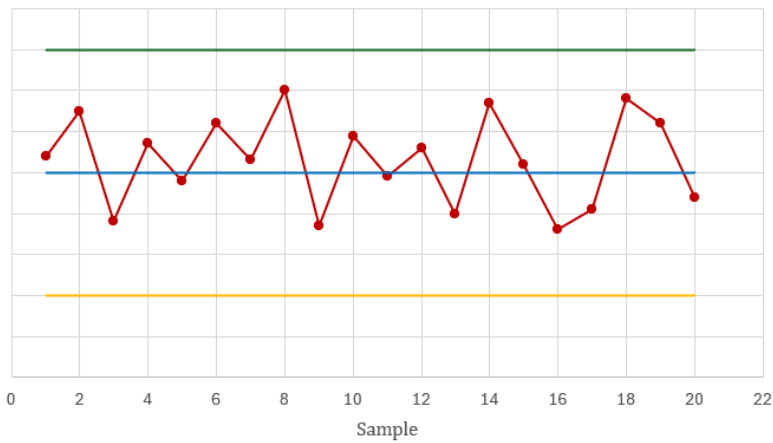
1. For each of the following control charts, determine if the process is in-control or out-of-control. If the process is out-of-control, identify the pattern in the data that indicates an out-of-control process.



a.

Click to see Answer

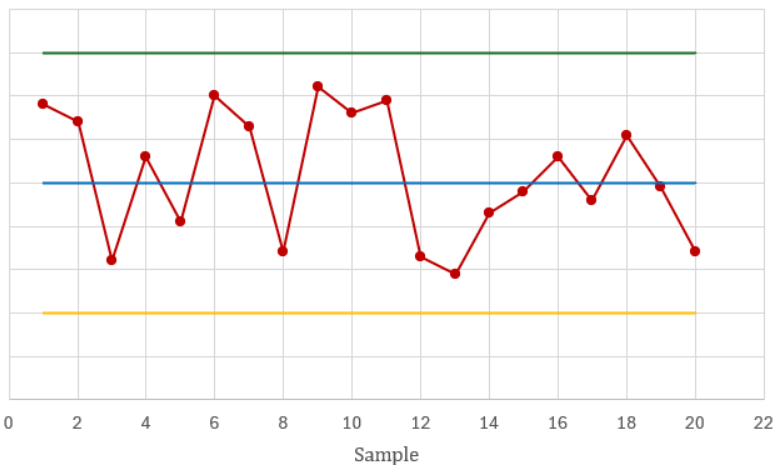
Out-of-control because there are points below the lower control limit.



b.

Click to see Answer

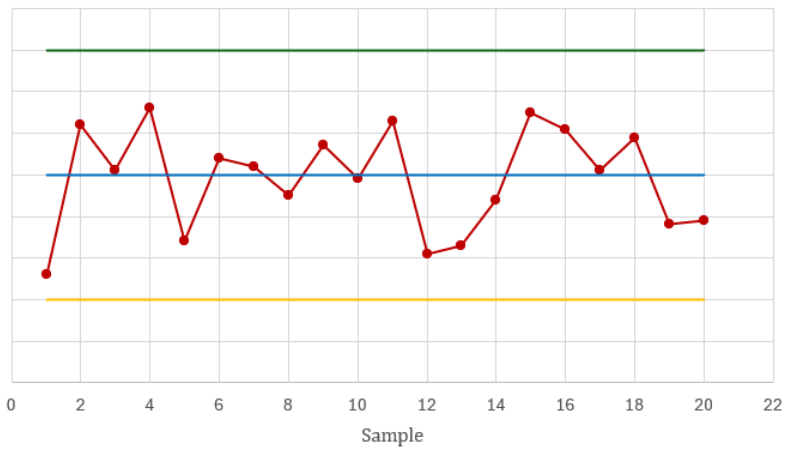
Out-of-control because there are 15 consecutive points that alternate in direction.



c.

Click to see Answer

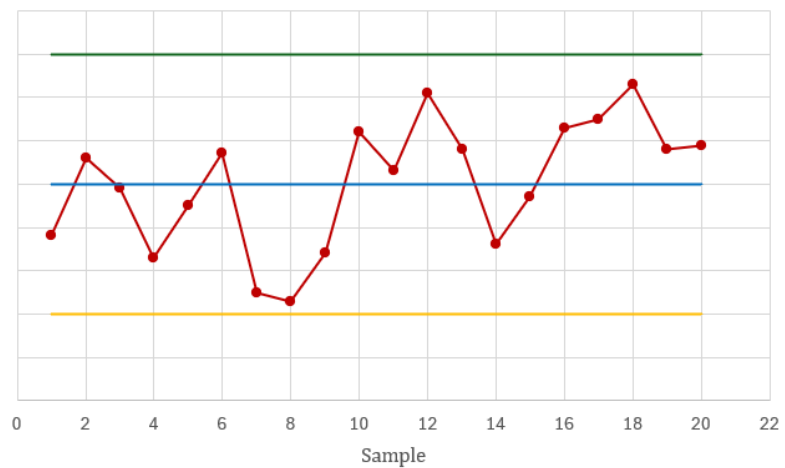
Out-of-control because there are 8 consecutive points, none within one standard deviation of the centerline.



d.

Click to see Answer

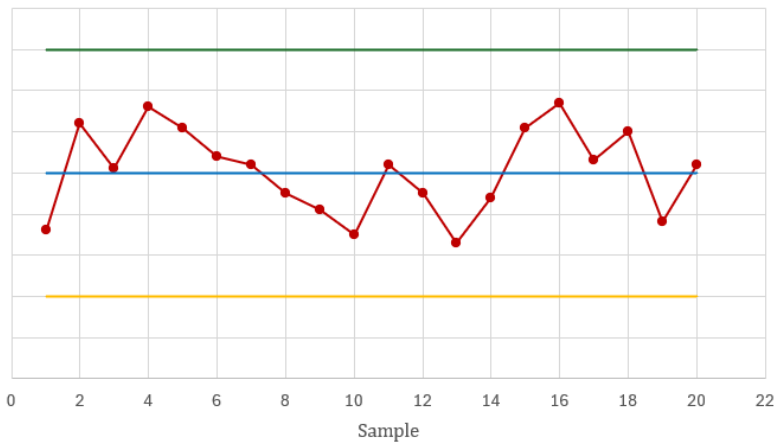
In-control.



e.

Click to see Answer

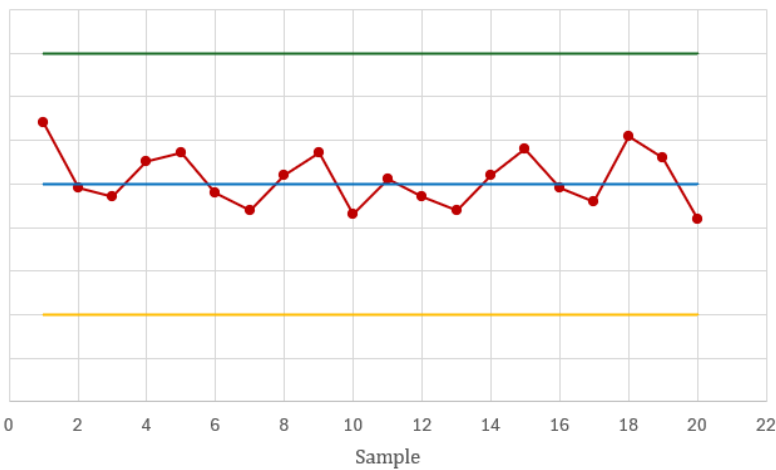
Out-of-control because there are two of three consecutive points more than two standard deviations from the centerline.



f.

Click to see Answer

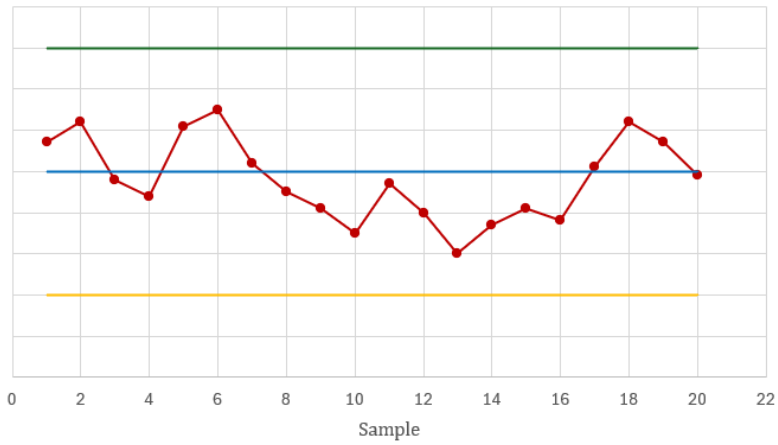
Out-of-control because there are 6 consecutive decreasing points.



g.

Click to see Answer

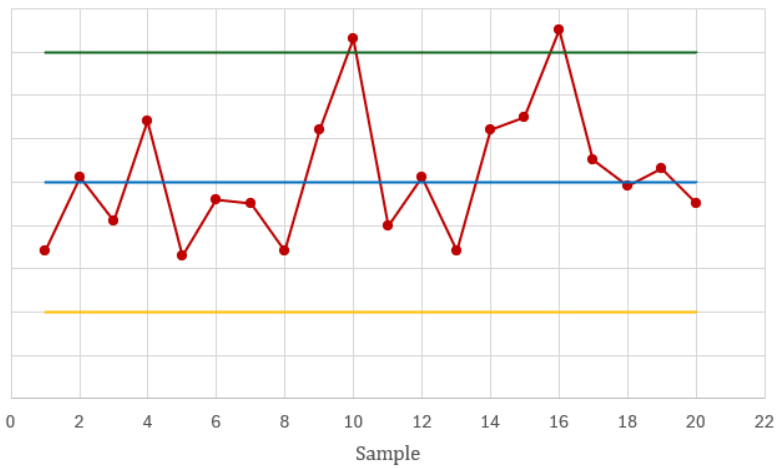
Out-of-control because there are 15 consecutive points, all within one standard deviation of the centerline.



h.

Click to see Answer

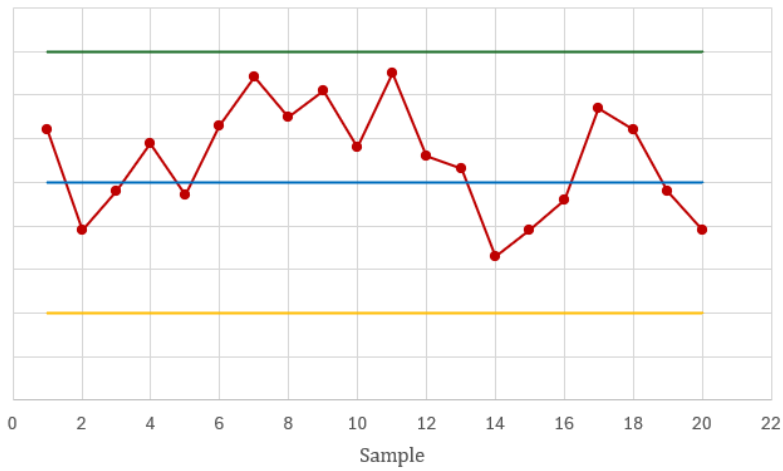
Out-of-control because there are 8 consecutive points all below the centerline.



i.

Click to see Answer

Out-of-control because there are points above the upper control limit.



j.

Click to see Answer

Out-of-control because there are four of five consecutive points more than one standard deviation from the centerline.

15.2 CONTROL CHARTS FOR VARIABLES

LEARNING OBJECTIVES

- Calculate the centerline, upper control limit, and lower control limit for an \bar{x} -chart or an R -chart.

Control charts that are used to monitor processes that are measured on a continuous scale are called **control charts for variables**. Examples of measurements that might require monitoring include weight, height, or volume. There are two different types of control charts to monitor measurements—control charts for means and control charts for range. Control charts for means, called **\bar{x} -charts**, monitor the mean or central tendency of the process. For example, if a company produces 300 gram bags of ground coffee, the \bar{x} -chart monitors the mean weight of the bags of coffee. Control charts for range, called **R -charts**, monitor the variability of the process. For example, if a company produces 300 gram bags of ground coffee, the R -chart monitors the range between the heaviest weight and the lightest weight of the bags of coffee. Generally, \bar{x} -charts and R -charts are used together in order to monitor both the average and variability of the process simultaneously.

\bar{x} -Charts

An \bar{x} -chart monitors the average or central tendency of a process. To construct an \bar{x} -chart, a collection of samples, all of the same size n , are taken from the process; the mean of each sample is calculated, and the sample means are plotted on the control chart. Because an \bar{x} -chart is based on samples taken from a population, \bar{x} -charts are based on the sampling distribution of the sample means and the Central Limit Theorem.

Recall that the sampling distribution of the sample means \bar{x} is the distribution of the sample means

from all possible samples of size n taken from a population. The Central Limit Theorem states that the distribution of the sample means follows a normal distribution if the population the samples are taken from is normal or if the sample size n is sufficiently large (i.e. $n \geq 30$). As well, the theorem states that

$$\mu_{\bar{x}} = \mu \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where $\mu_{\bar{x}}$ is the mean of the sample means, $\sigma_{\bar{x}}$ is the standard deviation of the sample means, μ is the population mean, σ is the population standard deviation, and n is the sample size.

Recall that the Empirical Rule for normal distributions states that 95% of the observations fall within two standard deviations of the mean and 99.7% of the observations fall within three standard deviations of the mean. Assuming the conditions of the Central Limit Theorem are met, the distribution of the sample means follows a normal distribution. So the Empirical Rule applies to the distribution of the sample means, and implies that:

- 95% of the sample means \bar{x} fall within two standard deviations $\sigma_{\bar{x}}$ of the mean $\mu_{\bar{x}}$. In other words, 95% of the sample means fall in between the values of $\mu_{\bar{x}} - 2\sigma_{\bar{x}}$ and $\mu_{\bar{x}} + 2\sigma_{\bar{x}}$.
- 99.7% of the sample means \bar{x} fall within three standard deviations $\sigma_{\bar{x}}$ of the mean $\mu_{\bar{x}}$. In other words, 99.7% of the sample means fall in between the values of $\mu_{\bar{x}} - 3\sigma_{\bar{x}}$ and $\mu_{\bar{x}} + 3\sigma_{\bar{x}}$.

This application of the Empirical Rule to the distribution of the sample means forms the basis for the upper and lower control limits on an \bar{x} -chart. In the context of quality control and an \bar{x} -chart, if the sample mean \bar{x} falls within three standard deviations above or below the mean value, then the process is considered in-control with 99.7% confidence. In other words, if a sample mean falls outside the three standard deviations, then the process is out-of-control with 99.7% confidence. Similarly, if the sample mean \bar{x} falls within two standard deviations above or below the mean value, then the process is considered in-control with 95% confidence.

Process Mean and Process Standard Deviation Known

The process mean is the population mean μ , and the process standard deviation is the population standard deviation σ . When both of these values are known, $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Then the centerline, the upper control limit, and the lower control limit for the \bar{x} -chart are:

$$\begin{array}{l}
 \text{Centerline} = \mu \\
 \text{Upper Control Limit} = \mu + z \times \frac{\sigma}{\sqrt{n}} \\
 \text{Lower Control Limit} = \mu - z \times \frac{\sigma}{\sqrt{n}} \\
 \text{process mean} = \mu \\
 \text{process standard deviation} = \sigma \\
 \text{sample size} = n \\
 \text{number of standard deviations} = z \\
 \text{(2 for 95\% confidence and 3 for 99.7\% confidence)}
 \end{array}$$

EXAMPLE

A company produces 300 gram bags of ground coffee. To monitor the average weight of the bags, 36 bags are selected every hour and weighed. Based on an analysis of old records, the standard deviation of the overall weight of the bags is estimated to be 8 grams.

1. At 99.7% confidence, calculate the centerline, the upper control limit, and the lower control limit for the \bar{x} -chart to monitor the average weight of the bags.
2. Suppose the sample means for samples taken over a 10-hour period are 303, 301, 302, 300, 294, 297, 299, 301, 298, 303. Construct the \bar{x} -chart. Is the mean of the process in-control? Explain.

Solution

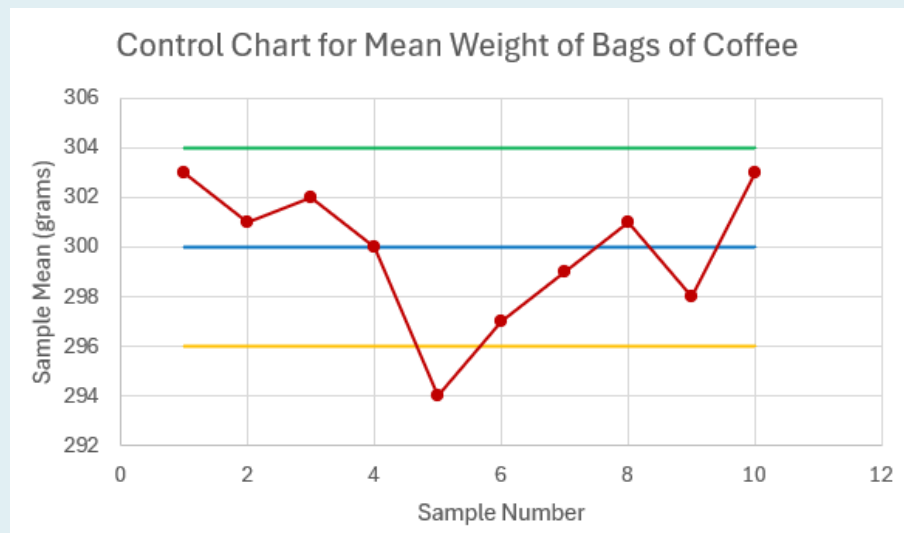
1. From the question, we have $\mu = 300$, $\sigma = 8$ and $n = 36$. At 99.7% confidence, $z = 3$.

$$\begin{aligned}\text{Centerline} &= \mu \\ &= 300 \text{ grams}\end{aligned}$$

$$\begin{aligned}\text{Upper Control Limit} &= \mu + z \times \frac{\sigma}{\sqrt{n}} \\ &= 300 + 3 \times \frac{8}{\sqrt{36}} \\ &= 304 \text{ grams}\end{aligned}$$

$$\begin{aligned}\text{Lower Control Limit} &= \mu - z \times \frac{\sigma}{\sqrt{n}} \\ &= 300 - 3 \times \frac{8}{\sqrt{36}} \\ &= 296 \text{ grams}\end{aligned}$$

2. The \bar{x} -chart is



The mean of the fifth sample falls below the lower control limit. This is strong evidence that the mean of the process is out-of-control.

NOTES

1. The sample means follow a normal distribution because the sample size **36** is greater than **30**.
2. Because this chart is for 99.7% confidence, the value of z in the formulas for the upper and lower control limits is **3**. This sets the upper and lower control limits at **3** standard deviations above and below the centerline, respectively.
3. The units for the centerline, upper control limit, and lower control limit are the same units as the data. In this example, the data is measured in grams, so the units of the centerline, upper control limit, and lower control limit are also in grams.

TRY IT

A company produces **30** cm long metal rods. To monitor the average length of the rods, **40** rods are selected every hour, and their length is measured. Based on an analysis of old records, the standard deviation of the overall length of the rods is estimated to be **0.5** cm. At 95% confidence, calculate the centerline, the upper control limit, and the lower control limit for the \bar{x} -chart to monitor the average length of the rods.

Click to see Solution

$$\begin{aligned}\text{Centerline} &= \mu \\ &= 30 \text{ cm}\end{aligned}$$

$$\begin{aligned}\text{Upper Control Limit} &= \mu + z \times \frac{\sigma}{\sqrt{n}} \\ &= 30 + 2 \times \frac{0.5}{\sqrt{40}} \\ &= 30.16 \text{ cm}\end{aligned}$$

$$\begin{aligned}\text{Lower Control Limit} &= \mu - z \times \frac{\sigma}{\sqrt{n}} \\ &= 30 - 2 \times \frac{0.5}{\sqrt{40}} \\ &= 29.84 \text{ cm}\end{aligned}$$

Process Mean Unknown and Process Standard Deviation Known

In many situations, the process mean is unknown. In these situations, the process mean μ is estimated using the mean of the sample means, called the **overall sample mean** $\bar{\bar{x}}$. In the process of constructing the \bar{x} -chart, samples are taken and their means calculated. The overall sample mean is just the mean of these sample means.

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k}$$

$$\begin{aligned}\bar{x}_j &= \text{mean of the } j\text{th sample} \\ k &= \text{number of samples}\end{aligned}$$

NOTE

The overall sample mean is just the mean of all of the data. Instead of calculating the individual

sample means and then calculating the mean of the sample means, the overall sample mean may be found by calculating the average of all of the collected data.

When the process mean is unknown, and the process standard deviation is known, the centerline, the upper control limit, and the lower control limit for the \bar{x} -chart are:

$$\begin{array}{l} \text{Centerline} = \overline{\overline{x}} \\ \text{Upper Control Limit} = \overline{\overline{x}} + z \times \frac{\sigma}{\sqrt{n}} \\ \text{Lower Control Limit} = \overline{\overline{x}} - z \times \frac{\sigma}{\sqrt{n}} \end{array}$$

$\overline{\overline{x}}$ = mean of the sample means
 σ = process standard deviation
 n = sample size
 z = number of standard deviations (2 for 95% confidence and 3 for 99.7% confidence)

EXAMPLE

A local beverage company bottles natural spring water. Each day the company takes six samples of 5 bottles each and measures the volume, in millimeters, of water in the bottles. The data is recorded in the table below. The standard deviation of the overall volume of water in the bottles is estimated to be 10 ml. Assume the volume of water in each bottle follows a normal distribution.

Sample	Volume per Bottle				
1	509.31	495.97	504.4	494.89	491.6
2	498.51	493.81	491.52	493	505.67
3	505.9	501.68	497.65	492.6	508.42
4	507.63	498.44	504.72	506.46	493.03
5	509.36	494.57	492.98	507.21	508.97
6	499.28	497.74	505.8	500.81	499.68

1. At 95% confidence, calculate the centerline, the upper control limit, and the lower control limit for the \bar{x} -chart to monitor the mean volume of water in the bottles.
2. Construct the \bar{x} -chart. Is the process in-control? Explain.

Solution

1. From the questions, we have $\sigma = 10$ and $n = 5$. Because the process mean μ is unknown, we need to calculate $\bar{\bar{x}}$. For each sample, calculate the sample mean.

Sample	Volume per Bottle					Sample Mean
1	509.31	495.97	504.4	494.89	491.6	499.234
2	498.51	493.81	491.52	493	505.67	496.502
3	505.9	501.68	497.65	492.6	508.42	501.25
4	507.63	498.44	504.72	506.46	493.03	502.056
5	509.36	494.57	492.98	507.21	508.97	502.618
6	499.28	497.74	505.8	500.81	499.68	500.662

Then, the mean of the sample means is

$$\begin{aligned}\bar{\bar{x}} &= \frac{499.234 + 496.502 + 501.25 + 502.056 + 502.618 + 500.662}{6} \\ &= 500.387 \text{ ml}\end{aligned}$$

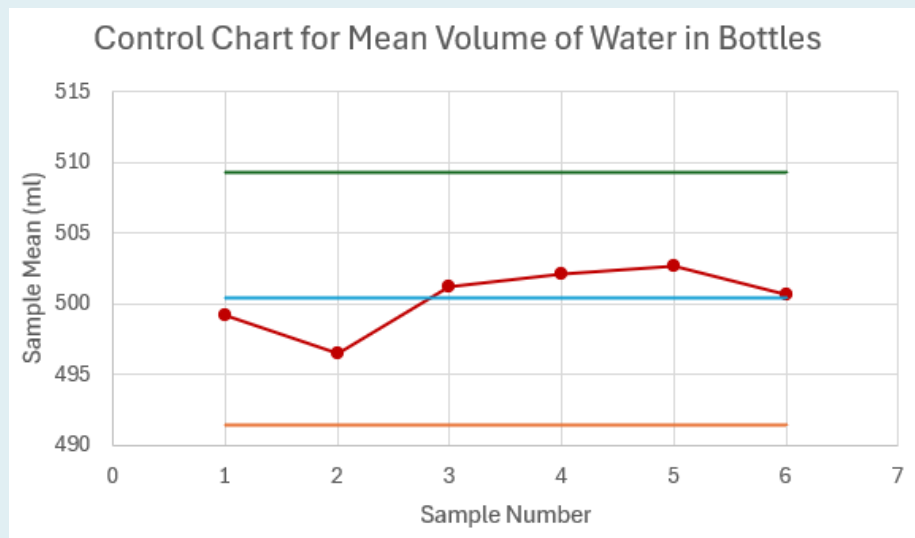
At 95% confidence, $z = 2$.

$$\begin{aligned}\text{Centerline} &= \bar{\bar{x}} \\ &= 500.387 \text{ ml}\end{aligned}$$

$$\begin{aligned}\text{Upper Control Limit} &= \bar{\bar{x}} + z \times \frac{\sigma}{\sqrt{n}} \\ &= 500.387 + 2 \times \frac{10}{\sqrt{5}} \\ &= 509.443 \text{ ml}\end{aligned}$$

$$\begin{aligned}\text{Lower Control Limit} &= \bar{\bar{x}} - z \times \frac{\sigma}{\sqrt{n}} \\ &= 500.387 - 2 \times \frac{10}{\sqrt{5}} \\ &= 491.443 \text{ ml}\end{aligned}$$

2. The \bar{x} -chart is



Based on the control chart, the mean of the process appears to be in-control.

NOTES

1. The sample means follow a normal distribution because the volume of water in the bottles follows a normal distribution.
2. Because this chart is for 95% confidence, the value of z in the formulas for the upper and lower control limits is **2**. This sets the upper and lower control limits at **2** standard deviations above and below the centerline, respectively.
3. Use Excel to perform the calculations on the raw data, instead of calculating the sample means and overall sample mean by hand.
4. The units for the centerline, upper control limit, and lower control limit in an \bar{x} -chart are the same units as the data. In this example, the data is measured in millilitres, so the units of the centerline, upper control limit, and lower control limit are also in millilitres.
5. On the \bar{x} -chart, we plot the sample mean for each sample, not the volume for each observation in the sample.

TRY IT

A manufacturer makes round shafts for drill presses. The manufacturer takes a series of samples of **10** shafts each and measures the diameter, in mm, of each shaft. The sample means for each sample are recorded in the table below. The standard deviation of the overall diameter of the shafts is estimated to be **2.5** mm. Assume the diameter of the shafts follows a normal distribution. Calculate the centerline, the upper control limit, and the lower control limit for the \bar{x} -chart at 99.7% confidence.

Sample	Sample Mean
1	141.11
2	138.23
3	140.69
4	141.76
5	139.73
6	139.4
7	139.65
8	142.42

Click to see Solution

$$\begin{aligned}\bar{\bar{x}} &= \frac{141.11 + 138.23 + 140.69 + 141.76 + 139.73 + 139.4 + 139.65 + 142.43}{8} \\ &= 140.37375 \text{ mm}\end{aligned}$$

$$\begin{aligned}\text{Centerline} &= \bar{\bar{x}} \\ &= 140.37375 \text{ mm}\end{aligned}$$

$$\begin{aligned}\text{Upper Control Limit} &= \bar{\bar{x}} + z \times \frac{\sigma}{\sqrt{n}} \\ &= 140.37375 + 3 \times \frac{2.5}{\sqrt{10}} \\ &= 142.75 \text{ mm}\end{aligned}$$

$$\begin{aligned}\text{Lower Control Limit} &= \bar{\bar{x}} - z \times \frac{\sigma}{\sqrt{n}} \\ &= 140.37375 - 3 \times \frac{2.5}{\sqrt{10}} \\ &= 138 \text{ mm}\end{aligned}$$

Process Mean and Process Standard Deviation Unknown

In many situations, both the process mean and the process standard deviation are unknown. As above, the process mean μ is estimated using the overall sample mean $\bar{\bar{x}}$. Although the average of the sample standard deviations could be used as an estimate of the process standard deviation, in practice, it is more common to use the average range of the samples in the calculation of the control limits for an \bar{x} -chart. Because range is easier to compute, the range can be used as an estimate of the process standard deviation in the \bar{x} -chart calculations with little computational effort.

In the process of constructing the \bar{x} -chart, samples are taken, and their means and ranges are calculated. The average of the sample ranges, \bar{R} , is found by averaging the ranges of the samples.

$$\bar{R} = \frac{R_1 + R_2 + \cdots + R_k}{k}$$

R_j = range of the j th sample

k = number of samples

When both the process mean, and the process standard deviation are unknown, the centerline, the upper control limit, and the lower control limit for the \bar{x} -chart are:

$$\text{Centerline} = \bar{\bar{x}}$$

$$\text{Upper Control Limit} = \bar{\bar{x}} + A_2 \times \bar{R}$$

$$\text{Lower Control Limit} = \bar{\bar{x}} - A_2 \times \bar{R}$$

$\bar{\bar{x}}$ = mean of the sample means

\bar{R} = mean of the sample ranges

A_2 = value from Factors for Control Limits Chart

Factors for Control Limits Chart			
Sample Size	A_2	D_3	D_4
2	1.880	0	3.267
3	1.023	0	2.575
4	0.729	0	2.282
5	0.577	0	2.115
6	0.483	0	2.004
7	0.419	0.076	1.924
8	0.373	0.136	1.864
9	0.337	0.184	1.816
10	0.308	0.223	1.777
11	0.285	0.256	1.744
12	0.266	0.284	1.716
13	0.249	0.308	1.692
14	0.235	0.329	1.671
15	0.223	0.348	1.652
16	0.212	0.364	1.636
17	0.203	0.379	1.621
18	0.194	0.392	1.608
19	0.187	0.404	1.596
20	0.180	0.414	1.586
21	0.173	0.425	1.575
22	0.167	0.434	1.566
23	0.162	0.443	1.557
24	0.157	0.452	1.548
25	0.153	0.459	1.541

NOTES

1. The value of A_2 in the calculation of the upper and lower control limits in an \bar{x} -chart is based on the sample size.
2. The Factors for Control Limits chart is based on 99.7% confidence.
3. The process standard deviation σ can be estimated by dividing the average range \bar{R} by a constant that depends on the sample size n . The value of A_2 results from substituting this estimate for σ into the formulas for the control limits with $z = 3$.

EXAMPLE

A company produces boxes of cereal. Each day, the company takes five samples of 4 boxes each and weighs the boxes, in grams, as part of the quality control process. The data is recorded in the table below. Assume the weight of the boxes follows a normal distribution. At 99.7% confidence, calculate the centerline, the upper control limit, and the lower control limit for the \bar{x} -chart to monitor the mean weight of the cereal boxes.

Sample	Weight of Boxes			
1	503.44	497.99	501.77	502.54
2	495.5	495.19	499.68	503.92
3	497.98	501.59	500.85	499.31
4	498.92	498.78	502.05	497.89
5	503.22	502.39	500.12	499.23

Solution

From the questions, we have $n = 4$. Because the process mean μ and the process standard

deviation are unknown, we need to calculate $\bar{\bar{x}}$ and \bar{R} . For each sample, calculate the sample mean and the range.

Sample	Weight of Boxes				Sample Mean	Sample Range
1	503.44	497.99	501.77	502.54	501.435	5.45
2	495.5	495.19	499.68	503.92	498.5725	8.73
3	497.98	501.59	500.85	499.31	499.9325	3.61
4	498.92	498.78	502.05	497.89	499.41	4.16
5	503.22	502.39	500.12	499.23	501.24	3.99

The mean of the sample means is

$$\begin{aligned}\bar{\bar{x}} &= \frac{501.435 + 498.5725 + 499.9325 + 499.41 + 501.24}{5} \\ &= 500.118 \text{ grams}\end{aligned}$$

The mean of the sample ranges is

$$\begin{aligned}\bar{R} &= \frac{5.45 + 8.73 + 3.61 + 4.16 + 3.99}{5} \\ &= 5.188 \text{ grams}\end{aligned}$$

On the Factors for Control Limits chart, the value for A_2 for a sample of size 4 is $A_2 = 0.729$.

$$\begin{aligned}\text{Centerline} &= \bar{\bar{x}} \\ &= 500.118 \text{ grams}\end{aligned}$$

$$\begin{aligned}\text{Upper Control Limit} &= \bar{\bar{x}} + A_2 \times \bar{R} \\ &= 500.118 + 0.729 \times 5.188 \\ &= 503.90 \text{ grams}\end{aligned}$$

$$\begin{aligned}\text{Lower Control Limit} &= \bar{\bar{x}} - A_2 \times \bar{R} \\ &= 500.118 - 0.729 \times 5.188 \\ &= 496.34 \text{ grams}\end{aligned}$$

NOTES

1. The sample means follow a normal distribution because the weight of the boxes follow a normal distribution.
2. Use Excel to perform the calculations on the raw data, instead of calculating the sample means, overall sample mean, sample ranges, and mean of the sample ranges by hand.
3. The units for the centerline, upper control limit, and lower control limit are the same units as the data. In this example, the data is measured in grams, so the units of the centerline, upper control limit, and lower control limit are also in grams.

TRY IT

Temperature is used to measure the output from a production process. The company takes 15 temperature readings, in Celsius, ten times a day as part of the quality control process. The sample means and sample ranges for each sample are recorded in the table below. Assume the temperatures follow a normal distribution. At 99.7\% confidence, calculate the centerline, the upper control limit, and the lower control limit for the \bar{x} -chart to monitor the mean temperature of the process.

Sample	Sample Mean	Sample Range
1	49.6	0.8
2	50.3	0.9
3	50.1	0.7
4	49.9	0.6
5	49.9	0.6
6	50.5	1
7	50.7	0.5
8	49.7	0.8
9	49.7	0.8
10	50.2	0.7

Click to see Solution

$$\begin{aligned}\bar{\bar{x}} &= \frac{49.6 + 50.3 + 50.1 + 49.9 + 49.9 + 50.5 + 50.7 + 49.7 + 49.7 + 50.2}{10} \\ &= 50.06^{\circ}\text{C}\end{aligned}$$

$$\begin{aligned}\bar{R} &= \frac{0.8 + 0.9 + 0.7 + 0.6 + 0.6 + 1 + 0.5 + 0.8 + 0.8 + 0.7}{10} \\ &= 0.74^{\circ}\text{C}\end{aligned}$$

$$\begin{aligned}\text{Centerline} &= \bar{\bar{x}} \\ &= 50.06^{\circ}\text{C}\end{aligned}$$

$$\begin{aligned}\text{Upper Control Limit} &= \bar{\bar{x}} + A_2 \times \bar{R} \\ &= 50.06 + 0.223 \times 0.74 \\ &= 50.23^{\circ}\text{C}\end{aligned}$$

$$\begin{aligned}\text{Lower Control Limit} &= \bar{\bar{x}} - A_2 \times \bar{R} \\ &= 50.06 - 0.223 \times 0.74 \\ &= 49.89^{\circ}\text{C}\end{aligned}$$

R-Charts

An \bar{x} -chart monitors the average or central tendency of a process. However, the variability of the process may be out-of-control, even when the average of the process is in-control. For example, suppose a company produces 300 gram bags of ground coffee. If the average of the filling process is in-control, the average of the bags of coffee is around 300 grams. But, if the variability is out-of-control, some bags may be radically underfilled and some bags may be radically overfilled, without changing the average of the process. So in addition to monitoring the average of the process with an \bar{x} -chart, a control chart for the ranges, called an R -chart, is used to monitor the variability of the process. To construct an R -chart, a collection of samples, all of the same size n , are taken from the process, the range of each sample is calculated, and the sample ranges are plotted on the control chart.

The centerline, the upper control limit, and the lower control limit for the R -chart are:

$$\text{Centerline} = \bar{R}$$

$$\text{Upper Control Limit} = D_4 \times \bar{R}$$

$$\text{Lower Control Limit} = D_3 \times \bar{R}$$

$$\bar{R} = \text{mean of the sample ranges}$$

$$D_3, D_4 = \text{values from Factors for Control Limits Chart}$$

NOTES

1. The values of D_3 and D_4 in the calculation of the upper and lower control limits in an R -chart are based on the sample size.
2. Because the Factors for Control Limits chart is based on 99.7\% confidence, all R -charts are at 99.7\% confidence.

EXAMPLE

A company produces boxes of cereal. Each day, the company takes five samples of 4 boxes each and weighs the boxes, in grams, as part of the quality control process. The data is recorded in the table below. Assume the weight of the boxes follows a normal distribution.

Sample	Weight of Boxes			
1	503.44	497.99	501.77	502.54
2	495.5	495.19	499.68	503.92
3	497.98	501.59	500.85	499.31
4	498.92	498.78	502.05	497.89
5	503.22	502.39	500.12	499.23

1. At 99.7\% confidence, calculate the centerline, the upper control limit, and the lower control limit for the R -chart to monitor the variability in the weight of the cereal boxes.
2. Construct the R -chart. Is the variability of the process in-control? Explain.

Solution

1. From the questions, we have $n = 4$. For each sample, calculate the sample range.

Sample	Weight of Boxes				Sample Range
1	503.44	497.99	501.77	502.54	5.45
2	495.5	495.19	499.68	503.92	8.73
3	497.98	501.59	500.85	499.31	3.61
4	498.92	498.78	502.05	497.89	4.16
5	503.22	502.39	500.12	499.23	3.99

The mean of the sample ranges is

$$\begin{aligned}\bar{R} &= \frac{5.45 + 8.73 + 3.61 + 4.16 + 3.99}{5} \\ &= 5.188 \text{ grams}\end{aligned}$$

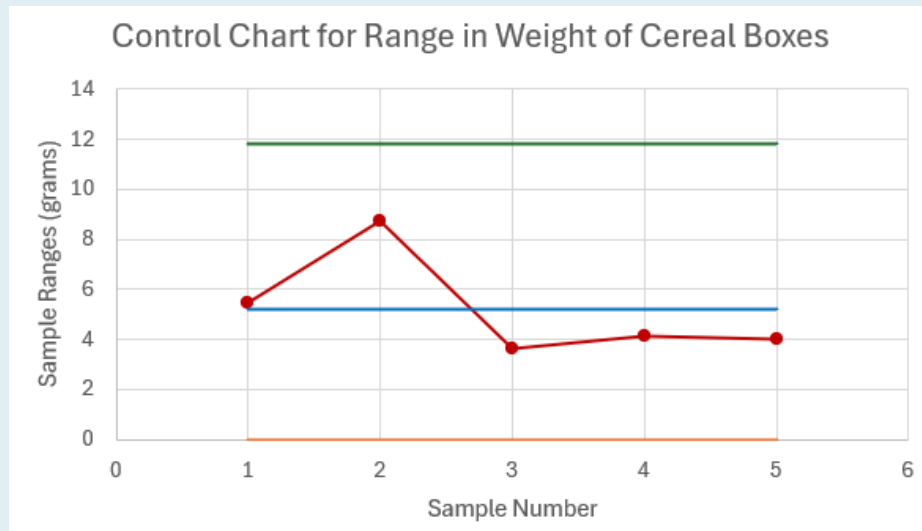
On the Factors for Control Limits chart, the values for D_3 and D_4 for a sample of size 4 are $D_3 = 0$ and $D_4 = 2.282$.

$$\begin{aligned}\text{Centerline} &= \bar{R} \\ &= 5.188 \text{ grams}\end{aligned}$$

$$\begin{aligned}\text{Upper Control Limit} &= D_4 \times \bar{R} \\ &= 2.282 \times 5.188 \\ &= 11.839 \text{ grams}\end{aligned}$$

$$\begin{aligned}\text{Lower Control Limit} &= D_3 \times \bar{R} \\ &= 0 \times 5.188 \\ &= 0 \text{ grams}\end{aligned}$$

2. The R -chart is



Based on the control chart, the variability of the process appears to be in-control.

NOTES

1. Use Excel to perform the calculations on the raw data, instead of calculating the sample ranges and mean of the sample ranges by hand.
2. The units for the centerline, upper control limit, and lower control limit for the R -chart are the same units as the data. In this example, the data is measured in grams, so the units of the centerline, upper control limit, and lower control limit are also in grams.
3. On the R -chart, we plot the sample ranges for each sample.

TRY IT

Temperature is used to measure the output from a production process. The company takes 15

temperature readings, in Celsius, ten times a day as part of the quality control process. The sample ranges for each sample are recorded in the table below. Assume the temperatures follow a normal distribution. At 99.7% confidence, calculate the centerline, the upper control limit, and the lower control limit for the \bar{R} -chart to monitor the variability in the temperature of the process.

Sample	Sample Range
1	0.8
2	0.9
3	0.7
4	0.6
5	0.6
6	1
7	0.5
8	0.8
9	0.8
10	0.7

Click to see Solution

$$\begin{aligned}\bar{R} &= \frac{0.8 + 0.9 + 0.7 + 0.6 + 0.6 + 1 + 0.5 + 0.8 + 0.8 + 0.7}{10} \\ &= 0.74^{\circ}\text{C}\end{aligned}$$

$$\begin{aligned}\text{Centerline} &= \bar{R} \\ &= 0.74^{\circ}\text{C}\end{aligned}$$

$$\begin{aligned}\text{Upper Control Limit} &= D_R \times \bar{R} \\ &= 1.652 \times 0.74 \\ &= 1.22248^{\circ}\text{C}\end{aligned}$$

$$\begin{aligned}\text{Lower Control Limit} &= D_3 \times \bar{R} \\ &= 0.348 \times 0.74 \\ &= 0.25752^{\circ}\text{C}\end{aligned}$$

Exercises

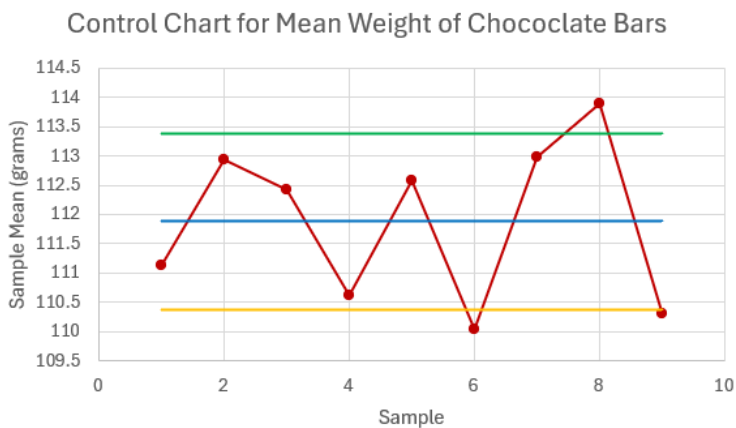
1. A company produces chocolate bars. The company wants to monitor the weight, in grams, of the bars. Each day, the company takes nine samples of 8 chocolate bars and weighs the bars in the sample. The sample means and sample ranges for one day's samples are recorded in the table below. Assume the weights of the bars follow a normal distribution.

Sample	Sample Mean	Sample Range
1	111.14	4.24
2	112.94	3.15
3	112.42	3.79
4	110.62	4.61
5	112.59	4.08
6	110.04	4.64
7	112.98	4.19
8	113.9	3.05
9	110.3	4.63

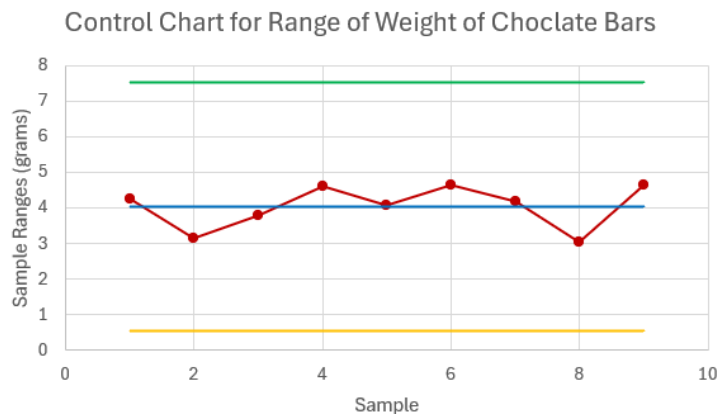
- Calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the average weight of the chocolate bars with 99.7% confidence.
- Calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the variability in the weight of the chocolate bars with 99.7% confidence.
- Construct the \bar{x} -chart and R -chart.
- Is the weight of the chocolate bars in-control? Explain.

Click to see Answer

- Centerline = 111.88 grams, Upper Control Limit = 113.89 grams, Lower Control Limit = 110.37 grams
- Centerline = 4.04 grams, Upper Control Limit = 7.53 grams, Lower Control Limit = 0.55 grams



c.



- The average weight of the chocolate bars is out-of-control because there are sample means outside of the control limits on the \bar{x} -chart. The variability of weight appears to be in-control on the R -chart.

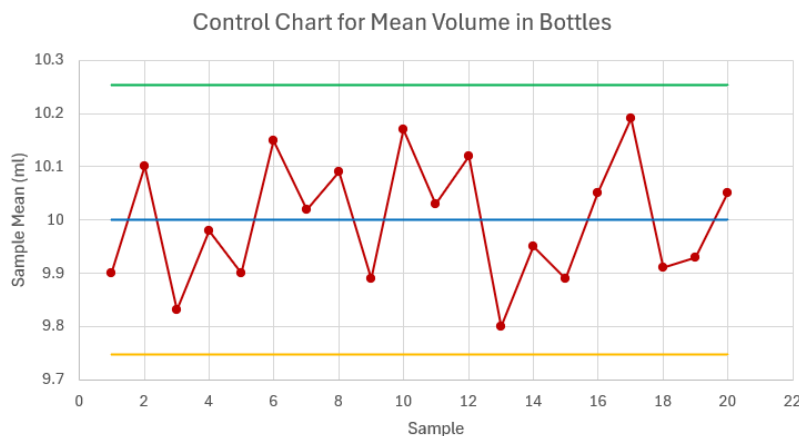
- A cosmetics company produces nail polish. The production process is set up to fill each bottle with 10 ml of nail polish. The standard deviation of the volume in the bottles is known to be

0.75 ml. The quality control department periodically selects samples of 35 bottles and measures the contents.

- Calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the average volume of nail polish in the bottles at 95% confidence.
- Explain what the control limits mean in the context of this question.
- The company takes twenty samples of 35 bottles. The sample means from the samples are 9.9, 10.1, 9.83, 9.98, 9.9, 10.15, 10.02, 10.09, 9.89, 10.17, 10.03, 10.12, 9.8, 9.95, 9.89, 10.05, 10.19, 9.91, 9.93, 10.05. Construct the \bar{x} -chart.
- Is the process in-control? Explain.

Click to see Answer

- Centerline = 10 ml, Upper Control Limit = 10.25 ml,
Lower Control Limit = 9.75 ml
- 95% of the sample means will fall between 9.75 ml and 10.25 ml.



- The process is out-of-control because there are fifteen consecutive points alternating in direction.

- A company manufactures household paint. During a 24-hour production cycle, the company took a sample of 12 paint cans every hour, and found the overall sample mean of volume in the cans was 3.9 litres and the mean of the sample ranges was 0.57 litres.

- Calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the mean volume of paint in the cans at 99.7% confidence.
- Calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the variability of the volume of paint in the cans at 99.7% confidence.

Click to see Answer

- Centerline = 3.9 liters, Upper Control Limit = 4.08 liters,
Lower Control Limit = 3.72 liters
- Centerline = 0.57 liters, Upper Control Limit = 0.98 liters,

Lower Control Limit = 0.16 liters

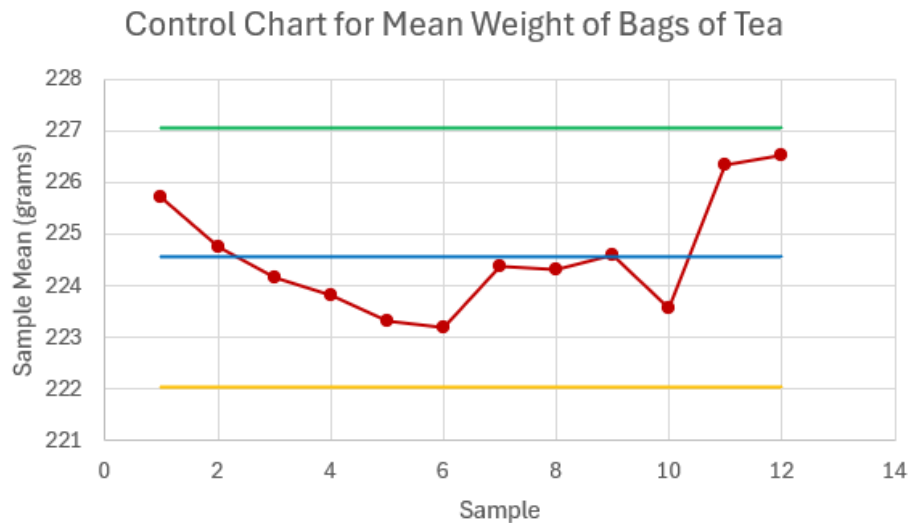
4. A tea manufacturer produces bags of loose-leaf tea that it sells to restaurants and coffee shops. As part of the quality control process, the manufacturer takes samples of the bags and measures the weight, in grams, of tea in the bags. The weights from the samples are recorded in the table below. From prior experience, the manufacturer knows that the standard deviation of the weight in the bags is 2.8 grams.

Sample	Weight of Bags of Tea				
1	220.92	228.74	224.47	226.74	227.73
2	226.66	227.74	221.94	226.64	220.8
3	224.99	224.18	221.77	226.02	223.85
4	223.17	222.65	221.32	225.63	226.24
5	225.41	222.22	221.42	223.55	223.92
6	222.79	224.98	223.11	221.35	223.78
7	220.65	225.09	224.19	225.76	226.13
8	221.05	223.49	225.85	229.85	221.28
9	227.05	228.59	223.95	220.01	223.32
10	223.35	221.34	223.58	227.31	222.27
11	225.28	227.86	221.88	229.99	226.63
12	225.84	226.31	222.77	228.61	229.17

- Calculate the centerline, upper control limit, and lower control limit for the \bar{x} -chart to monitor the weight of tea in the bags at 95% confidence.
- Construct the \bar{x} -chart.
- Is the weight of the bags of tea in-control? Explain.

Click to see Answer

- Centerline = 224.55 grams, Upper Control Limit = 227.06 grams,
Lower Control Limit = 222.05 grams



b.

- c. The weight of the bags of tea is out-of-control because there is a trend of six decreasing points on the control chart.

5. A company produces bags of pretzels. To ensure the bags have the proper weight, samples of 49 bags are taken and each bag is weighed. The overall mean of the samples is 225.17 grams. The standard deviation of the weights of the bags is 25 grams. At 99.7% confidence, calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the mean weight of the bags.

Click to see Answer

Centerline = 225.17 grams,

Upper Control Limit = 235.88 grams,

Lower Control Limit = 214.46 grams

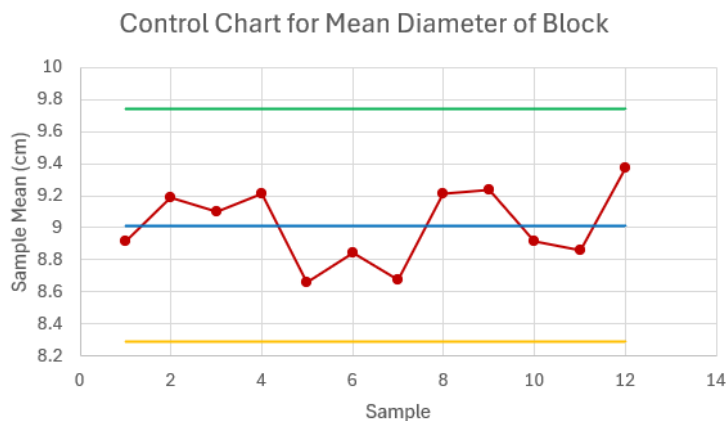
6. A company produces cylindrical blocks for a children's toy. The company needs to monitor the diameter, in centimetres, of the block to ensure the block fits into the toy. The company takes a series of samples and measures the diameter of blocks in each sample. The data is recorded in the table below. Assume the diameter of the blocks follows a normal distribution.

Sample	Diameter of Blocks					
1	9.1	8.66	9.27	8.12	8.4	9.94
2	9.9	9.08	8.31	9.49	8.93	9.44
3	9.87	8.62	8.76	9.71	8.23	9.4
4	9.48	8.79	9.85	8.26	9.27	9.63
5	8.21	9.76	8.16	8.48	8.26	9.06
6	8.67	9.2	8.32	8.94	8.06	9.87
7	8.18	8.48	8.36	9.73	9.17	8.12
8	9.42	9.38	9.37	8.17	9.87	9.05
9	9.81	8.78	9.18	8.9	9.69	9.06
10	9.52	8.6	9.38	8.6	9.06	8.34
11	8.57	8.3	9.55	8.93	8.53	9.27
12	9.57	9.41	9.98	9.4	8.75	9.12

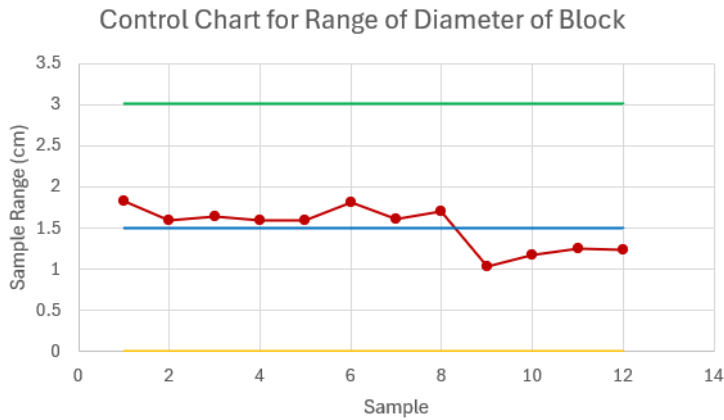
- Calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the mean diameter of the block with 99.7% confidence.
- Calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the variability in the diameter of the block with 99.7% confidence.
- Construct the \bar{x} -chart and R -chart.
- Is the diameter of the blocks in-control? Explain.

Click to see Answer

- Centerline = 9.015 cm, Upper Control Limit = 9.743 cm,
Lower Control Limit = 8.289 cm
- Centerline = 1.504 cm, Upper Control Limit = 3.014 cm,
Lower Control Limit = 0 cm



C.



- d. The average diameter of the blocks appears to be in-control on the \bar{x} -chart. The variability of the diameter appears to be out-of-control on the R -chart because there are eight consecutive points above the centerline.

7. Bags of chocolate candy are sampled to ensure proper weight. Ten samples are taken and the weight, in kg, of each bag is recorded in the table below.

Sample	Weight of Bags of Candy							
1	0.927	0.959	1.074	1.011	0.93	0.952	1.013	1.059
2	0.95	0.947	1.073	1.09	1.025	0.983	1.05	1.09
3	1.095	0.975	0.917	0.941	1.056	0.971	0.947	1.005
4	0.967	1.073	1.086	0.93	1.073	0.967	1.029	0.998
5	0.943	0.988	1.007	1.055	1.031	1.013	0.96	0.91
6	0.932	0.912	1.04	1.084	0.927	1.044	0.938	0.933
7	1.074	0.962	0.92	1.063	0.905	0.911	0.9	0.991
8	0.994	1.049	0.957	0.943	0.962	1.015	1.095	1.028
9	1.072	1.022	1.076	1.02	0.998	0.955	1.019	1.024
10	0.969	0.946	1.019	1.078	1.043	0.931	0.965	0.932

- Calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the mean weight of the bags with 99.7% confidence.
- Calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the variability in the weight of the bags with 99.7% confidence.

Click to see Answer

- a. Centerline = 0.9965 kg, Upper Control Limit = 1.0537 kg,
Lower Control Limit = 0.9392 kg
- b. Centerline = 0.1535 kg, Upper Control Limit = 0.2861 kg,
Lower Control Limit = 0.0209 kg

15.3 CONTROL CHARTS FOR ATTRIBUTES

LEARNING OBJECTIVES

- Calculate the centerline, upper control limit, and lower control limit for a p -chart or a c -chart.

Control charts that are used to monitor the percent of defective items or the number of defects are called **control charts for attributes**. There are two different types of control charts for attributes—control charts for the percent defective and control charts for the number of defects. Control charts for percent defective, called **p -charts**, monitor the percent of defective items in a sample. For example, if a company produces batteries, a p -chart might be used to monitor the percent of defective batteries produced each day. Control charts for number of defects, called **c -charts**, monitor the number of defects per item in a sample. For example, if a company produces lightbulbs, a c -chart might be used to monitor the number of defects per lightbulb in a sample of lightbulbs.

p -Charts

A p -chart monitors the percent of defective items in a sample. To construct a p -chart, a collection of samples, all of the same size n , are taken, the percent of defective items in each sample is calculated, and the sample proportions are plotted on the control chart. Because a p -chart is based on samples taken from a population, p -charts are based on the sampling distribution of the sample proportions.

Recall that the sampling distribution of the sample proportions \hat{p} is the distribution of the sample proportions from all possible samples of size n taken from a population. The distribution of the sample proportions follows a normal distribution if both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$. As well,

$$\mu_{\hat{p}} = p \quad \sigma_{\hat{p}} = \sqrt{\frac{p \times (1 - p)}{n}}$$

where $\mu_{\hat{p}}$ is the mean of the sample proportions, $\sigma_{\hat{p}}$ is the standard deviation of the sample proportions, p is the population proportion, and n is the sample size.

Recall that the Empirical Rule for normal distributions states that 95\% of the observations fall within two standard deviations of the mean and 99.7\% of the observations fall within three standard deviations of the mean. Assuming the $n \times p \geq 5$ and $n \times (1 - p) \geq 5$ conditions are met, the distribution of the sample proportions follows a normal distribution. The Empirical Rule applies to the distribution of the sample proportions and implies that:

- 95\% of the sample proportions \hat{p} fall within two standard deviations $\sigma_{\hat{p}}$ of the mean $\mu_{\hat{p}}$. In other words, 95\% of the sample proportions fall in between the values of $\mu_{\hat{p}} - 2\sigma_{\hat{p}}$ and $\mu_{\hat{p}} + 2\sigma_{\hat{p}}$.
- 99.7\% of the sample proportions \hat{p} fall within three standard deviations $\sigma_{\hat{p}}$ of the mean $\mu_{\hat{p}}$. In other words, 99.7\% of the sample proportions fall in between the values of $\mu_{\hat{p}} - 3\sigma_{\hat{p}}$ and $\mu_{\hat{p}} + 3\sigma_{\hat{p}}$.

This application of the Empirical Rule to the distribution of the sample proportions forms the basis for the upper and lower control limits on a p -chart. In the context of quality control and a p -chart, if the sample proportion \hat{p} falls within three standard deviations above or below the mean value, then the process is considered in-control with 99.7\% confidence. In other words, if a sample proportion falls outside the three standard deviations, then the process is out-of-control with 99.7\% confidence. Similarly, if the sample proportion \hat{p} falls within two standard deviations above or below the mean value, then the process is considered in-control with 95\% confidence.

The centerline, the upper control limit, and the lower control limit for the p -chart are:

$$\begin{array}{l} \text{Centerline} = \overline{p} \\ \text{Upper Control Limit} = \overline{p} + z \times \sqrt{\frac{\overline{p} \times (1 - \overline{p})}{n}} \\ \text{Lower Control Limit} = \overline{p} - z \times \sqrt{\frac{\overline{p} \times (1 - \overline{p})}{n}} \end{array}$$

\overline{p} = mean proportion of the population or mean of the sample proportions
 n = sample size
 z = number of standard deviations
 2 for 95\% confidence and 3 for 99.7\% confidence

NOTE

The value of \bar{p} depends on the information available. In the population proportion p is available, then \bar{p} is the value of p . In many situations, the population proportion is not available. In these cases, \bar{p} is the mean of the sample proportions.

EXAMPLE

A company manufactures batteries. In the past, the percent of defective batteries produced was 2.5%. Suppose samples of size 240 are taken. Calculate the centerline, the upper control limit, and the lower control limit for the p -chart to monitor the percent of defect batteries at 95% confidence.

Solution

From the question, we have $\bar{p} = 0.025$ and $n = 240$. At 95% confidence, $z = 2$.

$$\begin{aligned}\text{Centerline} &= \bar{p} \\ &= 0.025\end{aligned}$$

$$\begin{aligned}\text{Upper Control Limit} &= \bar{p} + z \times \sqrt{\frac{\bar{p} \times (1 - \bar{p})}{n}} \\ &= 0.025 + 2 \times \sqrt{\frac{0.025 \times (1 - 0.025)}{240}} \\ &= 0.0452\end{aligned}$$

$$\begin{aligned}\text{Lower Control Limit} &= \bar{p} - z \times \sqrt{\frac{\bar{p} \times (1 - \bar{p})}{n}} \\ &= 0.025 - 2 \times \sqrt{\frac{0.025 \times (1 - 0.025)}{240}} \\ &= 0.0048\end{aligned}$$

NOTES

1. The sample proportions follow a normal distribution because
 $n \times p = 240 \times 0.025 = 6 \geq 5$ and
 $n \times (1 - p) = 240 \times (1 - 0.025) = 234 \geq 5$.
2. Because this chart is for 95% confidence, the value of z in the formulas for the upper and lower control limits is 2. This sets the upper and lower control limits at 2 standard deviations above and below the centerline, respectively.
3. Because a p -chart monitors percents, the centerline, the upper control limit, and the lower control limit are percents. In this example, the centerline, upper control limit, and lower control limit are 2.5%, 4.52%, and 0.48%, respectively.

EXAMPLE

A company produces widgets. A quality control inspector monitors the percentage of defective widgets produced during the manufacturing process. For a 12-day period, the inspector takes a sample of 50 widgets a day and counts the number of defective widgets in the sample. The data is recorded in the table below.

Day	Number of Defective Widgets
1	9
2	8
3	6
4	9
5	10
6	4
7	1
8	2
9	8
10	4
11	6
12	7

1. Calculate the centerline, the upper control limit, and the lower control limit for the p -chart to monitor the percent of defective widgets at 99.7\% confidence.
2. Construct the p -chart. Is the process in-control? Explain.

Solution

1. From the question, we have $n = 50$. Because the population proportion p is unknown, we need to calculate the mean of the sample proportions to use for \bar{p} .

Day	Number of Defective Widgets	Sample Proportion
1	9	0.18
2	8	0.16
3	6	0.12
4	9	0.18
5	10	0.2
6	4	0.08
7	1	0.02
8	2	0.04
9	8	0.16
10	4	0.08
11	6	0.12
12	7	0.14

The mean of the sample proportions is

$$\bar{p} = \frac{0.18 + 0.16 + 0.12 + 0.18 + 0.2 + 0.08 + 0.02 + 0.04 + 0.16 + 0.08 + 0.12 + 0.14}{12}$$

$$= 0.1233 \dots$$

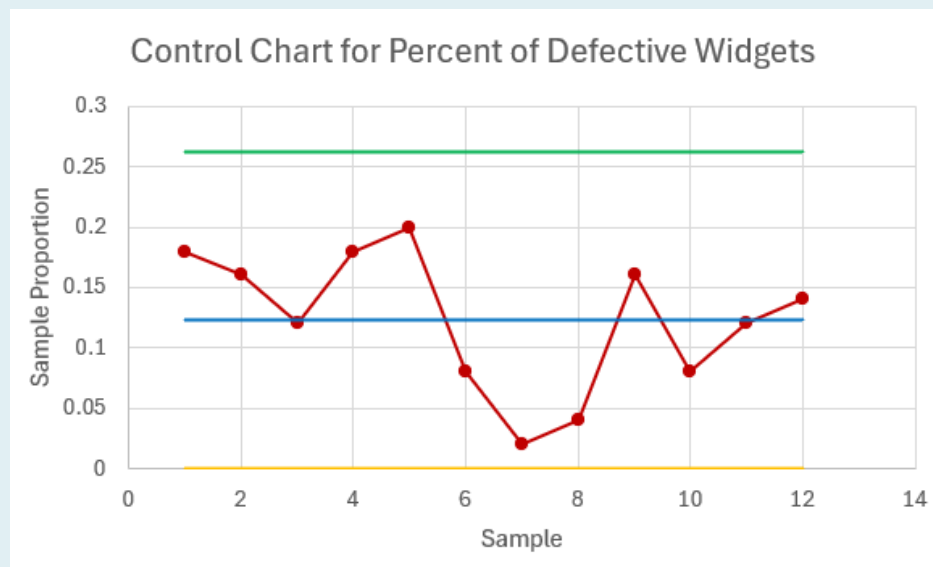
At 99.7\% confidence $z = 3$.

$$\begin{aligned}\text{Centerline} &= \bar{p} \\ &= 0.1233\end{aligned}$$

$$\begin{aligned}\text{Upper Control Limit} &= \bar{p} + z \times \sqrt{\frac{\bar{p} \times (1 - \bar{p})}{n}} \\ &= 0.1233 \dots + 3 \times \sqrt{\frac{0.1233 \dots \times (1 - 0.1233 \dots)}{50}} \\ &= 0.2628\end{aligned}$$

$$\begin{aligned}\text{Lower Control Limit} &= \bar{p} - z \times \sqrt{\frac{\bar{p} \times (1 - \bar{p})}{n}} \\ &= 0.1233 \dots - 3 \times \sqrt{\frac{0.1233 \dots \times (1 - 0.1233 \dots)}{50}} \\ &= -0.0162 \rightarrow 0\end{aligned}$$

2. The p -chart is



Based on the control chart, the percent of defective widgets appears to be in-control.

NOTES

1. The sample proportions follow a normal distribution because
 $n \times \bar{p} = 50 \times 0.1233 \dots = 6.166 \dots \geq 5$ and
 $n \times (1 - p) = 50 \times (1 - 0.1233 \dots) = 43.833 \dots \geq 5$.
2. To calculate each sample proportion, divide the number of defective widgets by the sample size. For example, the sample proportion for day 1 is $\hat{p} = \frac{9}{50} = 0.18$.
3. Because this chart is for 99.7% confidence, the value of z in the formulas for the upper and lower control limits is **3**. This sets the upper and lower control limits at **3** standard deviations above and below the centerline, respectively.
4. When the formula for the lower control limit in a p -chart produces a negative number, the lower control limit is set to **0**. Because the p -chart monitors the percent defective, it does not make sense to have a negative lower control limit, and so the convention in such cases is to make the lower control limit equal to **0**.
5. Use Excel to perform the calculations on the raw data, instead of calculating the sample proportions and mean of the sample proportions by hand.
6. Because a p -chart monitors percents, the centerline, the upper control limit, and the lower control limit are percents. In this example, the centerline, upper control limit, and lower control limit are 12.33%, 26.28%, and 0%, respectively.
7. On the p -chart, we plot the sample proportion for each sample, not the number of defective items. A p -chart monitors the percent defective, not the number of defective items.

TRY IT

A restaurant monitors the proportion of customer complaints per day. Over a 10-day period, the restaurant takes a sample of 80 customers and records the proportion of customer complaints in the

sample. The data is recorded in the table below. Calculate the centerline, the upper control limit, and the lower control limit for the control chart to monitor the percent of customer complaints per day with 99.7% confidence.

Day	Proportion of Complaints
1	0.1
2	0.15
3	0.1125
4	0.025
5	0.075
6	0.125
7	0.0125
8	0.0125
9	0.075
10	0.05

Click to see Solution

$$\begin{aligned}\bar{p} &= \frac{0.1 + 0.15 + 0.1125 + 0.025 + 0.075 + 0.125 + 0.0125 + 0.0125 + 0.075 + 0.05}{10} \\ &= 0.07375\end{aligned}$$

$$\begin{aligned}\text{Centerline} &= \bar{p} \\ &= 0.07375\end{aligned}$$

$$\begin{aligned}\text{Upper Control Limit} &= \bar{p} + z \times \sqrt{\frac{\bar{p} \times (1 - \bar{p})}{n}} \\ &= 0.07375 + 3 \times \sqrt{\frac{0.07375 \times (1 - 0.07375)}{80}} \\ &= 0.1614\end{aligned}$$

$$\begin{aligned}\text{Lower Control Limit} &= \bar{p} - z \times \sqrt{\frac{\bar{p} \times (1 - \bar{p})}{n}} \\ &= 0.07375 - 3 \times \sqrt{\frac{0.07375 \times (1 - 0.07375)}{80}} \\ &= -0.0139 \rightarrow 0\end{aligned}$$

c-Charts

A c -chart monitors the number of defects per item in a sample. To construct a c -chart, a single sample is taken, the number of defects on each item in the sample is counted, and the numbers of defects per item are plotted on the control chart. Because a c -chart is based on counting the number of defects that occur in a fixed sample or time interval, a c -chart is based on a Poisson distribution.

Recall that in a Poisson distribution, λ is the average number of occurrences in an interval, the mean of a Poisson distribution is λ , and the standard deviation is $\sqrt{\lambda}$. In a c -chart, \bar{c} is the average number of defects per item, which corresponds to the value of λ in the Poisson distribution, and the standard deviation equals $\sqrt{\bar{c}}$.

The control limits for a control chart are set to two or three standard deviations from the centerline. Consequently, the centerline, the upper control limit, and the lower control limit for the c -chart are:

$$\begin{array}{l} \text{Centerline} = \bar{c} \\ \text{Upper Control Limit} = \bar{c} + z \times \sqrt{\bar{c}} \\ \text{Lower Control Limit} = \bar{c} - z \times \sqrt{\bar{c}} \end{array}$$

where \bar{c} = mean number of occurrences per item, z = number of standard deviations (2 for 95% confidence and 3 for 99.7% confidence)

EXAMPLE

An online retailer wants to monitor the number of visits to their website each day. Over a 14-day period, the total number of visits to the website was 1,827. Calculate the centerline, the upper control limit, and the lower control limit for the \bar{c} -chart to monitor the number of visits to the website each day at 95% confidence.

Solution

From the question, we have $\bar{c} = \frac{1827}{14} = 130.5$ visits per day. At 95% confidence, $z = 2$.

$$\begin{aligned} \text{Centerline} &= \bar{c} \\ &= 130.5 \text{ visits per day} \end{aligned}$$

$$\begin{aligned} \text{Upper Control Limit} &= \bar{c} + z \times \sqrt{\bar{c}} \\ &= 130.5 + 2 \times \sqrt{130.5} \\ &= 164.77 \text{ visits per day} \end{aligned}$$

$$\begin{aligned} \text{Lower Control Limit} &= \bar{c} - z \times \sqrt{\bar{c}} \\ &= 130.5 - 2 \times \sqrt{130.5} \\ &= 96.23 \text{ visits per day} \end{aligned}$$

NOTES

1. The value of \bar{c} is the average number of visits per day to the website from the sample. In this case, the **1,872** is the total over the 14-day time period. To find the average per day, we need to divide the total by the number of days.
2. Because this chart is for 95% confidence, the value of z in the formulas for the upper and lower control limits is **2**. This sets the upper and lower control limits at **2** standard deviations above and below the centerline, respectively.

EXAMPLE

A company produces metal parts. A quality control inspector monitors the number of defects in each part. Each day, the inspector takes a sample of parts and counts the number of defects on each part. The data for yesterday's sample is recorded in the table below.

Item	Number of Defects
1	0
2	5
3	3
4	4
5	4
6	2
7	3
8	0
9	4
10	3
11	4
12	0
13	1
14	2

1. Calculate the centerline, the upper control limit, and the lower control limit for the \bar{c} -chart to monitor the number of defects per part at 99.7% confidence.
2. Construct the \bar{c} -chart. Is the process in-control? Explain.

Solution

1. First, calculate the mean of the number of defects per part.

$$\begin{aligned}\bar{c} &= \frac{0 + 5 + 3 + 4 + 4 + 2 + 3 + 0 + 4 + 3 + 4 + 0 + 1 + 2}{14} \\ &= 2.5 \text{ defects per part}\end{aligned}$$

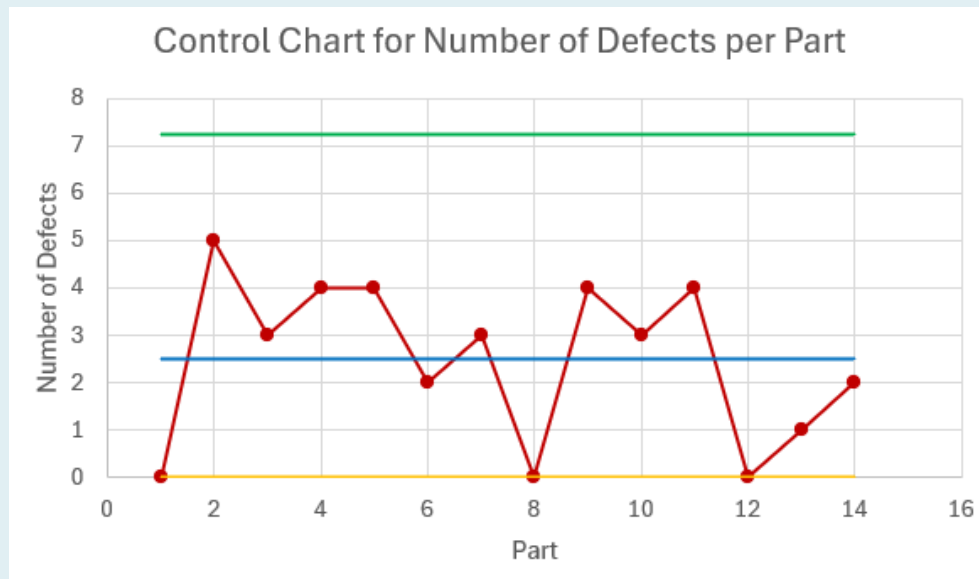
At 99.7% confidence $z = 3$.

$$\begin{aligned}\text{Centerline} &= \bar{c} \\ &= 2.5 \text{ defects per part}\end{aligned}$$

$$\begin{aligned}\text{Upper Control Limit} &= \bar{c} + z \times \sqrt{\bar{c}} \\ &= 2.5 + 3 \times \sqrt{2.5} \\ &= 7.24 \text{ defects per part}\end{aligned}$$

$$\begin{aligned}\text{Lower Control Limit} &= \bar{c} - z \times \sqrt{\bar{c}} \\ &= 2.5 - 3 \times \sqrt{2.5} \\ &= -2.24 \rightarrow 0 \text{ defects per part}\end{aligned}$$

2. The c -chart is



Based on the control chart, the number of defects per part appears to be in-control.

NOTES

1. Because this chart is for 99.7% confidence, the value of z in the formulas for the upper and lower control limits is **3**. This sets the upper and lower control limits at **3** standard deviations above and below the centerline, respectively.

2. When the formula for the lower control limit in a c -chart produces a negative number, the lower control limit is set to 0. Because the c -chart monitors the number of defects, it does not make sense to have a negative lower control limit, and so the convention in such cases is to make the lower control limit equal to 0.
3. Use Excel to find the average of the number of defects column.
4. On the c -chart, we plot the number of defects per item.

TRY IT

A company produces five products per hour. The company wants to track the number of defects in each hour's batch of product. Over the course of a 10-hour production cycle, the number of defects each hour was recorded in the table below. Calculate the centerline, the upper control limit, and the lower control limit for the control chart to monitor the number of defects per hour at 99.7% confidence.

Hour	Number of Defects
1	2
2	1
3	3
4	0
5	3
6	2
7	1
8	2
9	3
10	2

Click to see Solution

$$\begin{aligned}\bar{c} &= \frac{2 + 1 + 3 + 0 + 3 + 3 + 2 + 1 + 2 + 3 + 2}{10} \\ &= 1.9 \text{ defects per hour}\end{aligned}$$

$$\begin{aligned}\text{Centerline} &= \bar{c} \\ &= 1.9 \text{ defects per hour}\end{aligned}$$

$$\begin{aligned}\text{Upper Control Limit} &= \bar{c} + z \times \sqrt{\bar{c}} \\ &= 1.9 + 3 \times \sqrt{1.9} \\ &= 6.035 \text{ defects per hour}\end{aligned}$$

$$\begin{aligned}\text{Lower Control Limit} &= \bar{c} - z \times \sqrt{\bar{c}} \\ &= 1.9 - 3 \times \sqrt{1.9} \\ &= -2.235 \rightarrow 0 \text{ defects per hour}\end{aligned}$$

Exercises

1. A company produces marble countertops. A defect in the countertop requires the entire countertop to be reconstructed. In a sample of 80 countertops, 3 were defective. Calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the percentage of defective countertops at 95% confidence.

Click to see Answer

$\text{Centerline} = 3.75\%$,

$\text{Upper Control Limit} = 7.998\%$,

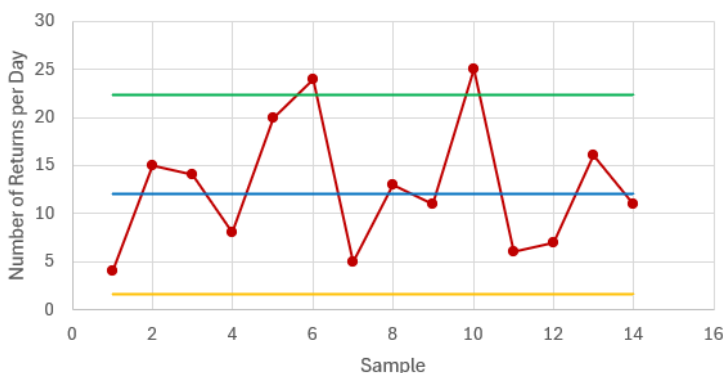
$\text{Lower Control Limit} = 0\%$

2. A retail store monitors the number of customer returns per day. Under normal conditions, the store expects 12 customer returns per day.
 - a. Calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the number of customer returns per day at 99.7% confidence.
 - b. Over the past two weeks, the store recorded the following number of customer returns per day: 4, 15, 14, 8, 20, 24, 5, 13, 11, 25, 6, 7, 16, 11. Construct the control chart to monitor the number of customer returns per day.
 - c. Is the process in-control? Explain.

Click to see Answer

- a. Centerline = 12 returns per day,
Upper Control Limit = 22.39 returns per day,
Lower Control Limit = 1.61 returns per day

Control Chart for Number of Returns per Day



- b.
 - c. The process is out-of-control because there are points above the upper control limit.
3. A company manufactures batteries. The quality control department monitors the percentage

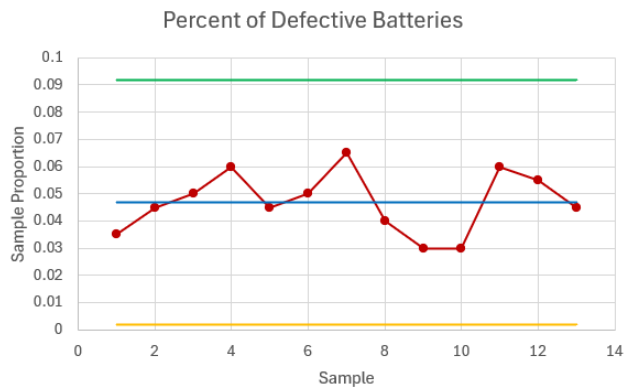
of defective batteries produced each day. A quality control inspector takes a sample of 200 batteries over a 13-day period and records the number of defective batteries in each sample. The data is recorded in the table below.

Day	Number of Defective Batteries
1	7
2	9
3	10
4	12
5	9
6	10
7	13
8	8
9	6
10	6
11	12
12	11
13	9

- Calculate the centerline, upper control limit, and lower control limit to monitor the percent of defective batteries made each day at 99.7\% confidence.
- Construct the control chart to monitor the percentage of defective batteries.
- Is the process in-control? Explain.

Click to see Answer

- $\text{Centerline} = 4.69\%$, $\text{Upper Control Limit} = 9.18\%$,
 $\text{Lower Control Limit} = 0.21\%$



b.

c. Based on the p -chart, the process appears in-control.

4. A retail store receives complaints about the attitude of some of its sales associates. Over a 10-day period, the store records the number of complaints they receive per day. The data is recorded in the table below. Calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the number of complaints per day at 95% confidence.

Day	Number of Complaints
1	22
2	23
3	25
4	24
5	28
6	23
7	20
8	27
9	25
10	23

Click to see Answer

Centerline = 24 complaints per day,

Upper Control Limit = 33.798 complaints per day,

Lower Control Limit = 14.202 complaints per day

5. A computer software manufacturer offers its customers a 24-hour helpline if they have problems or questions about the software. In one 24-hour period, the company received 384 calls to its helpline. Calculate the centerline, upper control limit, and lower control limit for the control chart to monitor the number of calls per hour to the helpline with 99.7% confidence.

Click to see Answer

Centerline = 16 calls per hour, Upper Control Limit = 28 calls per hour,
Lower Control Limit = 4 calls per hour

6. A manufacturer of flash drives wants to monitor the percentage of defective drives produced. Over a 10-day period, the manufacturer takes a sample of 160 drives each day and tests each drive. The proportion of defective drives in each sample is recorded in the table below.

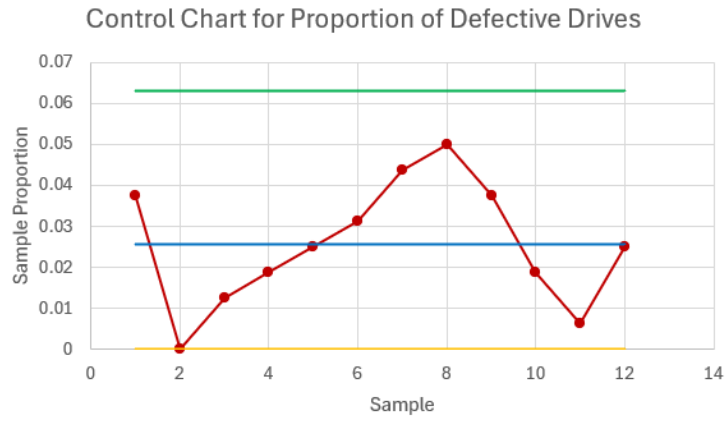
Sample	Proportion Defective
1	0.0375
2	0
3	0.0125
4	0.01875
5	0.025
6	0.03125
7	0.04375
8	0.05
9	0.0375
10	0.01875
11	0.00625
12	0.025

- Calculate the centerline, upper control limit, and lower control limit to monitor the proportion of defective drives made each day at 99.7% confidence.
- Construct the control chart to monitor the proportion of defective batteries.
- Is the process in-control? Explain.

Click to see Answer

- $\text{Centerline} = 2.55\%$, $\text{Upper Control Limit} = 6.29\%$,

\text{Lower Control Limit}=0\%



b.

- c. Based on the p -chart, the process appears out-of-control because there are six consecutive increasing points.

REFERENCES

1.1 Definitions of Statistics, Probability, and Key Terms

The Data and Story Library. (n.d.). Retrieved May 1, 2013, from <http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html>

1.2 Types of Data and Levels of Measurement & 1.3 Sampling and Sampling Techniques

Book of Odds. (n.d.). *How George Gallup Picked the President*. <http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President>

Gallup. (n.d.). *Gallup Presidential Election Trial-Heat Trends, 1936–2008*. Retrieved May 1, 2013, from <http://www.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4>

Gallup-Healthways Well-Being Index. (n.d.). Retrieved May 1, 2013, from <http://www.well-beingindex.com/default.asp>

Gallup-Healthways Well-Being Index. (n.d.). Retrieved May 1, 2013, from <http://www.well-beingindex.com/methodology.asp>

Gallup-Healthways Well-Being Index. (n.d.) Retrieved May 1, 2013, from <http://www.gallup.com/poll/146822/gallup-healthways-index-questions.aspx>

LBCC Distance Learning (DL) program data in 2010-2011. (n.d.). Retrieved May 1, 2013, from <http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus>

Lusinchi, D. (2012) “‘President’ Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?” *Social Science History* ,36(1)1, 23-54. <http://ssh.dukejournals.org/content/36/1/23.abstract>

San Jose Mercury News. (n.d.).

The Literary Digest Poll,” Virtual Laboratories in Probability and Statistics. (n.d.). Retrieved May 1, 2013, from <http://www.math.uah.edu/stat/data/LiteraryDigest.html>

The Data and Story Library. (n.d.). Retrieved May 1, 2013, from <http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html>

Lane, D. 2003, June 20. *Levels of Measurement.* OpenStax CNX. Retrieved May 1, 2013, from <http://cnx.org/content/m10809/latest/>

Levels of Measurement. (n.d.). Retrieved May 1, 2013, from http://infinity.cos.edu/faculty/woodbury/stats/tutorial/Data_Levels.htm

State & County QuickFacts. (n.d.). Retrieved May 1, 2013, from http://quickfacts.census.gov/qfd/download_data.html

State & County QuickFacts: Quick, easy access to facts about people, business, and geography. (n.d.). U.S. Census Bureau. Retrieved from May 1, 2013, <http://quickfacts.census.gov/qfd/index.html>

Table 5: Direct hits by mainland United States Hurricanes (1851-2004). (n.d.). National Hurricane Center. Retrieved May 1, 2013, from <http://www.nhc.noaa.gov/gifs/table5.gif>

Taylor, C. 2018, February 2. *The Levels of Measurement in Statistics.* Thoughtco. <http://statistics.about.com/od/HelpandTutorials/a/Levels-Of-Measurement.htm>

1.4 Experimental Design and Ethics

Alden, L. (2013, May 1). *Statistics can be Misleading.* econoclass.com. Retrieved May 1, 2013, from <http://www.econoclass.com/misleadingstats.html>

America’s Best Small Companies. (n.d.). Forbes. Retrieved May 1, 2013, from, <http://www.forbes.com/best-small-companies/list/>

April 2013 Air Travel Consumer Report. (2013, April 11). U.S. Department of Transportation. Retrieved, May 1, 2013, from, <http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report> (accessed May 1, 2013).

Bhattacharjee, Y. (2013, April 26). The Mind of a Con Man. *The New York Times Magazine.* http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?src=dayp&_r=2&

- Data. (n.d.). BusinessWeek. Retrieved May 1, 2013, from <https://www.bloomberg.com/businessweek>
- Data. (n.d.). Forbes. Retrieved May 1, 2013, from, <https://www.forbes.com/>
- Earthquake Information by Year*. (n.d.). U.S. Geological Survey. Retrieved, May 1, 2013, from <http://earthquake.usgs.gov/earthquakes/eqarchives/year/>
- Jacson, M. L., Croft, R. J., Kennedy, G. A., Owens, K., & Howard, M. E. (2013). *Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors*. *Accident Analysis and Prevention*, *Jan*(50), 438-44. <http://www.ncbi.nlm.nih.gov/pubmed/22721550>
- Levelt, W. J. M., Drenth, P., & Noort, E. (Eds.). (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Tilburg: Commissioned by the Tilburg University, University of Amsterdam and the University of Groningen. <http://hdl.handle.net/11858/00-001M-0000-0010-258A-9>
- McClung, M., & Collins, D. (2007). "Because I know it will!": Placebo effects of an ergogenic aid on athletic performance. *Journal of Sport & Exercise Psychology*, *29*(3), 382-94. <https://doi.org/10.1123/jsep.29.3.382>
- Medina, de los A. (2007, November 19). *Ethics in Statistics*. OpenStax CNX. Retrieved May 1, 2013, from <http://cnx.org/contents/12a5d87b-3fe3-4606-b368-cef865e40cde@1>
- Mehta, A. (2011, July 21). *Daily Dose of Aspiring Helps Reduce Heart Attacks: Study*. International Business Times. Retrieved May 1, 2013, from, <http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443>
- National Highway Traffic Safety Administration. (n.d.). *Fatality Analysis Report Systems (FARS) Encyclopedia*. Retrieved May 1, 2013, from <http://www-fars.nhtsa.dot.gov/Main/index.aspx>
- Nutrition Source: Vitamin E*. (n.d.). Harvard T.H. Chan School of Public Health. Retrieved May 1, 2013, from <http://www.hsph.harvard.edu/nutritionsource/vitamin-e/>
- Reents, S. (2008, February 4). *Don't Underestimate the Power of Suggestion*. AthleteInMe.com. Retrieved May 1, 2013, from <http://www.athleteinme.com/ArticleView.aspx?id=1053>
- The Data and Story Library* (n.d.). Retrieved May 1, 2013, from <http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html>
- U.S. Department of Health and Human Services. (2019). Code of Federal Regulations: Title 45

Public Welfare Department of Health and Human Services, Part 46 Protection of Human Subjects, Section 46.111: Criteria for IRB Approval of Research.

2.1 Histograms, Frequency Polygons, and Time Series Graphs

Births Time Series Data. (2013). General Register Office For Scotland. Retrieved April 3, 2013, from, <http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html>

CO2 emissions (kt). (2013). The World Bank. Retrieved April 3, 2013, from, <http://databank.worldbank.org/data/home.aspx>

Consumer Price Index. (n.d.). United States Department of Labor: Bureau of Labor Statistics. Retrieved April 3, 2013, from, <http://data.bls.gov/pdq/SurveyOutputServlet>

Demographics: Children under the age of 5 years underweight. (n.d.). Indexmundi. Retrieved April 3, 2013, from, <http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en>

Food Security Statistics. (n.d.). Food and Agriculture Organization of the United Nations. Retrieved April 3, 2013, from, <http://www.fao.org/economic/ess/ess-fs/en/>

Gunst, R., & Mason, R. (1980). *Regression Analysis and Its Application: A Data-Oriented Approach.* CRC Press.

Overweight and Obesity: Adult Obesity Facts. (n.d.). Centers for Disease Control and Prevention. Retrieved April 3, 2013, from, <http://www.cdc.gov/obesity/data/adult.html>

Presidents. (2007). Fact Monster. Retrieved April 3, 2013, from, <http://www.factmonster.com/ipka/A0194030.html>

Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more. (2013). Scholastic. Retrieved April 3, 2013, from, <http://www.scholastic.com/teachers/article/timeline-guide-us-presidents>

2.2 Measures of Central Tendency

Data. (n.d.). The World Bank. Retrieved April 3, 2013, from, <http://www.worldbank.org>

Demographics: Obesity – adult prevalence rate. (n.d.). Indexmundi. Retrieved April 3, 2013, from, <http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en>

2.4 Measures of Position

1990 Census. (n.d.). United States Department of Commerce: United States Census Bureau. Retrieved April 3, 2013, from, <http://www.census.gov/main/www/cen1990.html>

Cauchon, D., & Overberg, P. (2012). Census data shows minorities now a majority of U.S. births. *USA Today*. Retrieved April 3, 2013, from, <http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1>

Data. (n.d.). *San Jose Mercury News*.

Data. (n.d.). The United States Department of Commerce: United States Census Bureau. Retrieved April 3, 2013, from, <http://www.census.gov/>

Yankelovich Partners. (n.d.). Survey. *Time Magazine*.

2.5 Measures of Variability

Data. (n.d.). In *Microsoft Bookshelf*.

King, B. (n.d.). *Graphically Speaking*. Institutional Research, Lake Tahoe Community College. Retrieved April 3, 2013, from, <http://www.ltcc.edu/web/about/institutional-research>

3.1 The Terminology of Probability

Worldatlas. (2013). Countries List by Continent. In *Worldatlas.com*. Retrieved May 2, 2013, from, <http://www.worldatlas.com/cntycont.htm>

3.2 Contingency Tables

Blood Types. (n.d.). American Red Cross. Retrieved May 3, 2013, from, <http://www.redcrossblood.org/learn-about-blood/blood-types>

Data. (n.d.). National Center for Health Statistics, The United States Department of Health and Human Services.

Data. (n.d.). United States Senate. Retrieved May 2, 2013, from, <https://www.senate.gov/>

Haiman, C. A., Stram, D. O., Wilkens, L. R., Pike, M. C., Kolonel, L. N., Henderson, B. E., & le Marchand, L. (2006, January 26). Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer. *The New England Journal of Medicine*. <http://www.nejm.org/doi/full/10.1056/NEJMoa033250>

Human Blood Types. (2011). Unite Blood Services. Retrieved May 2, 2013, from, <http://www.unitedbloodservices.org/learnMore.aspx>

Samuel, T. M. (2013). *Strange Facts about RH Negative Blood*. eHow Health. Retrieved May 2, 2013, from, http://www.ehow.com/facts_5552003_strange-rh-negative-blood.html

United States: Uniform Crime Report – State Statistics from 1960–2011. (n.d.). The Disaster Center. Retrieved May 2, 2013, from, <http://www.disastercenter.com/crime/>

3.6 Joint Probabilities

Data. (n.d.). *Baseball-Almanac*. Retrieved May 2, 2013, from, <https://www.baseball-almanac.com/>

Data. (n.d.). Field Research Corporation.

Data. (n.d.). The Roper Center: Public Opinion Archives at the University of Connecticut. Retrieved May 2, 2013, from, <http://www.ropercenter.uconn.edu/>

Data. (n.d.). The Wall Street Journal. <https://www.wsj.com/>

Data. (n.d.). U.S. Census Bureau. <https://www.census.gov/>

DiCamillo, M., & Field, M. (n.d.). *The File Poll*. Field Research Corporation. Retrieved May 2, 2013, from, <http://www.field.com/fieldpollonline/subscribers/Rls2443.pdf>

Mayor's Approval Down. (n.d.). Forum Research. Retrieved May 2, 2013, from, http://www.forumresearch.com/forms/News_Archives/News_Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf

Rider, D. (2011, September 14). Ford support plummeting, poll suggests. *The Star*. Retrieved May 2, 2013, from, http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html

Roulette. (n.d.). In *Wikipedia*. <http://en.wikipedia.org/wiki/Roulette>

Shin, H. B., & Kominski, R. A. (2010, April 1). *Language Use in the United States: 2007*. United States Census Bureau. <https://www.census.gov/library/publications/2010/acs/acs-12.html>

4.3 Expected Value and Standard Deviation of a Discrete Random Variable

Course Catalog. (n.d.). Florida State University. Retrieved May 15, 2013, from, https://m.fsu.edu/default/course_catalog/index

World Earthquakes: Live Earthquake News and Highlights. (2012). World Earthquakes. Retrieved May 15, 2013, from, http://www.world-earthquakes.com/index.php?option=ethq_prediction

4.4 The Binomial Distribution

Access to electricity (% of population). (2013). The World Bank. Retrieved May 15, 2015, from, http://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?order=wbapi_data_value_2009%20wbapi_data_value%20wbapi_data_value-first&sort=asc

Distance Education. (n.d.). In *Wikipedia*. Retrieved May 15, 2013, from, http://en.wikipedia.org/wiki/Distance_education

NBA Statistics – 2013. ESPN. Retrieved May 15, 2013, from, http://espn.go.com/nba/statistics/_/seasontype/2

Newport, F. (2013, May 9). *Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income*. Gallup. <http://www.gallup.com/poll/162368/americans-enjoy-saving-rather-spending.aspx>

Pryor, J. H., DeAngelo, L., Palucki Blake, L., Hurtado, S., & Tran, S. (2011). *The American Freshman: National Norms Fall 2011*. Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA. <http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf>

The World FactBook. (n.d.). Central Intelligence Agency. Retrieved May 15, 2013, from, <https://www.cia.gov/library/publications/the-world-factbook/geos/af.html>

What are the key statistics about pancreatic cancer? (2013). American Cancer Society. Retrieved May

15, 2013, from, <http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics>

4.5 The Poisson Distribution

ATL Fact Sheet. (2013). Department of Aviation at the Hartsfield-Jackson Atlanta International Airport. Retrieved February 18, 2019, from, <http://www.atl.com/about-atl/atl-factsheet/>

Children and Childrearing. (n.d.). Ministry of Health, Labour, and Welfare. Retrieved May 15, 2013, from, <http://www.mhlw.go.jp/english/policy/children/children-childrearing/index.html>

Daily Mail Reporter. (2011, June 9). One born every minute: The maternity unit where mothers are THREE to a bed. *Daily Mail*. Retrieved May 15, 2013, from, <http://www.dailymail.co.uk/news/article-2001422/Busiest-maternity-ward-planet-averages-60-babies-day-mothers-bed.html>

Eating Disorder Statistics. (2006). South Carolina Department of Mental Health. Retrieved May 15, 2013, from, <http://www.state.sc.us/dmh/anorexia/statistics.htm>

Giving Birth in Manila. (2011, June 8). *The Guardian*. Retrieved May 15, 2013, from, <http://www.theguardian.com/world/gallery/2011/jun/08/philippines-health#/?picture=375471900&index=2>

Lenhart, A. (2012, March 19). *Teens, Smartphones & Texting*. Pew Research Center. Retrieved May 15, 2013, from, http://www.pewinternet.org/~media/Files/Reports/2012/PIP_Teens_Smartphones_and_Texting.pdf

Smith, A. (2011, September 19). *How Americans Use Text Messaging*. Pew Research Center. Retrieved May 15, 2013, from, <http://pewinternet.org/Reports/2011/Cell-Phone-Texting-2011/Main-Report.aspx>

Teen Drivers: Fact Sheet, Injury Prevention & Control: Motor Vehicle Safety. (2012, October 2). Center for Disease Control and Prevention. Retrieved May 15, 2013, from, http://www.cdc.gov/Motorvehiclesafety/Teen_Drivers/teendrivers_factsheet.html

Vanderkam, L. (2012, October 8). *Stop Checking Your Email, Now*. Fortune. Retrieved May 15, 2013, from, <http://management.fortune.cnn.com/2012/10/08/stop-checking-your-email-now/>

World Earthquakes: Live Earthquake News and Highlights. (n.d.). World Earthquakes Live. Retrieved May 15, 2013, from, http://www.world-earthquakes.com/index.php?option=ethq_prediction

5.3 The Standard Normal Distribution

2012 College-Bound Seniors Total Group Profile Report. (2012). CollegeBoard. Retrieved May 14, 2013, from, <http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf>

Blood Pressure of Males and Females. (n.d.). StatCrunch. Retrieved May 14, 2013, from, <http://www.statcrunch.com/5.0/viewreport.php?reportid=11960>

Data. (n.d.). National Basketball Association. Retrieved May 14, 2013, from, www.nba.com

Data. (n.d.). *San Jose Mercury News.*

Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009. (2009). National Center for Education Statistics. Retrieved May 14, 2013, from, http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp

Janssen, S. (Ed.). (n.d.). *The World Almanac and Book of Facts.* World Almanac Books.

List of stadiums by capacity. (n.d.). In *Wikipedia*. Retrieved May 14, 2013, from, https://en.wikipedia.org/wiki/List_of_stadiums_by_capacity

The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores. (2009). London School of Hygiene and Tropical Medicine. Retrieved May 14, 2013, from, http://conflict.lshtm.ac.uk/page_125.htm

5.4 Calculating Probabilities for a Normal Distribution

Facebook Statistics. (n.d.). Statistics Brain. Retrieved May 14, 2013, from, <http://www.statisticbrain.com/facebook-statistics/>

Naegele's rule. (n.d.). In *Wikipedia*. Retrieved May 14, 2013, from, http://en.wikipedia.org/wiki/Naegele's_rule

NUMMI. (2010, March 26). *This American Life*. Retrieved May 14, 2013, from, <http://www.thisamericanlife.org/radio-archives/episode/403/nummi>

Scratch-Off Lottery Ticket Playing Tips. (n.d.). WinAtTheLottery.com. Retrieved May 14, 2013, from, <http://www.winatthelottery.com/public/departement40.cfm>

Smart Phone Users, By The Numbers. (n.d.). Visual.ly. Retrieved May 14, 2013, from, <http://visual.ly/smart-phone-users-numbers>

6.1 Sampling Distribution of the Sample Mean

Baran, D. (n.d.). *20 Percent of Americans Have Never Used Email.* WebGuild. Retrieved May 14, 2013, from, <http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email>

Data. (n.d.). The Flurry Blog. Retrieved May 17, 2013, from, <http://blog.flurry.com>

Data. (n.d.). The United States Department of Agriculture.

7.2 Confidence Intervals for a Single Population Mean with Known Population Standard Deviation

American Fact Finder. (n.d.). U.S. Census Bureau. Retrieved July 2, 2013, from, <http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t>

Disclosure Data Catalog: Candidate Summary Report 2012. (n.d.). U.S. Federal Election Commission. Retrieved July 2, 2013, from, <http://www.fec.gov/data/index.jsp>

Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall. (n.d.). Foothill De Anza Community College District. Retrieved September 30, 2013, from, http://research.fhda.edu/factbook/FH_Demo_Trends/FoothillDemographicTrends.htm

Kuczmarski, R. J., Ogden, C. L., Guo, S. S., Grummer-Strawn, L. M., Flegal, K. M., Mei, Z. , Wei, R., Curtin, L. R., Roche, A. F., & Johnson, C. L. (2002, May). Vital Health Statistics: 2000 CDC Growth Charts for the United States: Methods and Development. *Centers for Disease Control and Prevention*, 11(246). Retrieved July 2, 2013, from, <http://www.cdc.gov/growthcharts/2000growthchart-us.pdf>

Mean Income in the Past 12 Months (in 2011 Inflation-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates. (n.d.). American Fact Finder, U.S. Census Bureau. Retrieved July 2, 2013, from, http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1902&prodType=table

Metadata Description of Candidate Summary File. (n.d.). U.S. Federal Election Commission.

Retrieved July 2, 2013, from, <http://www.fec.gov/finance/disclosure/metadata/metadataforcandidatesummary.shtml>

National Health and Nutrition Examination Survey. (n.d.). Centers for Disease Control and Prevention. Retrieved July 2, 2013, from, <http://www.cdc.gov/nchs/nhanes.htm>

Ralph, N., & German, K. (2011, June 1). *Cell phones with the highest radiation levels (pictures).* CNET. Retrieved July 2, 2013, from, <http://reviews.cnet.com/cell-phone-radiation-levels/>

7.3 Confidence Intervals for a Single Population Mean with Unknown Population Standard Deviation

America's Best Small Companies. (2013). Forbes. Retrieved July 2, 2013, from, <http://www.forbes.com/best-small-companies/list/>

Data. (n.d.). Businessweek. <http://www.businessweek.com/>.

Data. (n.d.). Forbes. <http://www.forbes.com/>.

Data. (n.d.). In *Microsoft Bookshelf*.

Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012. (n.d.). Federal Election Commission. Retrieved July 2, 2013, from, <http://www.fec.gov/data/index.jsp>

Human Toxome Project: Mapping the Pollution in People. (n.d.). Environmental Working Group. Retrieved July 2, 2013, from, <http://www.ewg.org/sites/humantoxome/participants/participant-group.php?group=in+utero%2Fnewborn>

Metadata Description of Leadership PAC List. (n.d.). Federal Election Commission. Retrieved July 2, 2013, from, <http://www.fec.gov/finance/disclosure/metadata/metadataLeadershipPacList.shtml>

7.4 Confidence Intervals for Population Proportions

2013 Teen and Privacy Management Survey. (n.d.). Pew Research Center: Internet and American Life Project. Retrieved July 2, 2013, from, http://www.pewinternet.org/~media/Files/Questionnaire/2013/Methods%20and%20Questions_Teens%20and%20Social%20Media.pdf

52% Say Big-Time College Athletics Corrupt Education Process. (2013, May 16). Rasmussen Reports.

Retrieved July 2, 2013, from, http://www.rasmussenreports.com/public_content/lifestyle/sports/may_2013/52_say_big_time_college_athletics_corrupt_education_process

Jensen, T. (2013, May 10). *Democrats, Republicans Divided on Opinion of Music Icons*. Public Policy Polling. Retrieved July 2, 2013, from, <https://www.publicpolicypolling.com/polls/democrats-republicans-divided-on-opinion-of-music-icons/>

Madden, M., Lenhart, A., Coresi, S., Gasser, U., Duggan, M., Smith, A., & Beaton, M. (2013, May 21). *Teens, Social Media, and Privacy*. Pew Research Center. Retrieved July 2, 2013, from, <https://www.pewresearch.org/internet/2013/05/21/teens-social-media-and-privacy/>

Saad, L. (2013, May 23). *Three in Four U.S. Workers Plan to Work Pas Retirement Age: Slightly more say they will do this by choice rather than necessity*. Gallup. Retrieved July 2, 2013, from, <http://www.gallup.com/poll/162758/three-four-workers-plan-work-past-retirement-age.aspx>

The Field Poll. (n.d.). Field. Retrieved July 2, 2013, from, <http://field.com/fieldpollonline/subscribers/>

Zogby. (2013, May 16). *New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure 'Investment' for National Security*. Zogby Analytics. Retrieved July 2, 2013, from, <http://www.zogbyanalytics.com/news/299-americans-neither-worried-nor-prepared-in-case-of-a-disaster-sunyit-zogby-analytics-poll>

8.1 Null and Alternative Hypotheses

Data. (n.d.). The National Institute of Mental Health. <http://www.nimh.nih.gov/publicat/depression.cfm>

8.4 Hypothesis Tests for a Population Mean with Known Population Standard Deviation, 8.5 Hypothesis Tests for a Population Mean with Unknown Population Standard Deviation, 8.6 Hypothesis Tests for a Population Proportion

Allen, E. I., & Seaman, J. (2005). *Growing by Degrees: Online Education in the United States, 2005*. The Sloan Consortium.

Amit Schitai, A. (n.d.). Data.

- Data. (n.d.). American Automobile Association. Retrieved June 27, 2013, from, www.aaa.com
- Data. (n.d.). American Library Association. Retrieved June 27, 2013, from, <https://www.ala.org/>
- Data. (n.d.). Bureau of Labor Statistics. <http://www.bls.gov/oes/current/oes291111.htm>.
- Data. (n.d.). Centers for Disease Control and Prevention. Retrieved June 27, 2013, from, www.cdc.gov
- Data. (n.d.). Energy.Gov. Retrieved June 27, 2013, from, <http://energy.gov>
- Data. (n.d.). Gallup. Retrieved June 27, 2013, from <https://www.gallup.com/home.aspx>
- Data. (n.d.). La Leche League International. <http://www.lalecheleague.org/Law/BAFeb01.html>
- Data. (n.d.). Toastmasters International. <http://toastmasters.org/artisan/detail.asp?CategoryID=1&SubCategoryID=10&ArticleID=429&Page=1>.
- Data. (n.d.). United States Census Bureau. Retrieved June 27, 2013, from, <https://www.census.gov/programs-surveys/sis/resources/data-tools/quickfacts.html>
- Data. (n.d.). United States Census Bureau. <http://www.census.gov/hhes/socdemo/language/>.
- Data, (n.d.). Weather Underground. Retrieved June 27, 2013, from, <https://www.wunderground.com/>
- Deprez, E. E. *NYC Smoking Rate Falls to Record Low of 14%, Bloomberg Says*. Businessweek. Retrieved June 27, 2013, from <https://www.bloomberg.com/news/articles/2011-09-15/new-york-city-adult-smoking-rate-falls-to-all-time-low-of-14-mayor-says#:~:text=New%20York's%20adult%20smoking%20rate,are%20smoking%2C%20the%20mayor%20said>
- FBI. (n.d.). *Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005*. The Disaster Center. Retrieved June 27, 2013, from, <http://www.disastercenter.com/kentucky/crime/3868.htm>
- Foothill-De Anza Community College District*. (2006, Winter). De Anza College. http://research.fhda.edu/factbook/DAdemofs/Fact_sheet_da_2006w.pdf
- Johansen, C., Boice, Jr., J., McLaughlin, J., Olsen, J. (2001). Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark. *Journal of National Cancer Institute*, 93(3), 203-207. <https://doi.org/10.1093/jnci/93.3.203>

How often does sexual assault occur? (n.d.). RAINN. Retrieved June 27, 2013, from, <http://www.rainn.org/get-information/statistics/frequency-of-sexual-assault>

9.1 Statistical Inference for Two Population Means with Known Population Standard Deviations

Data. (n.d.). United States Census Bureau. <http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf>

FBI. (n.d.). *Texas Crime Rates 1960–1012*. Uniform Crime Reports, The Disaster Center. Retrieved June 17, 2013, from, <http://www.disastercenter.com/crime/txcrime.htm>

Hinduja, S. (2013). *Sexting Research and Gender Differences*. Cyberbullying Research Center. Retrieved June 17, 2013, from, <http://cyberbullying.us/blog/sexting-research-and-gender-differences/>

Smart Phone Users, By the Numbers. (2013). Visually. Retrieved June 17, 2013, from, <http://visual.ly/smart-phone-users-numbers>

Smith, A. (2013). *35% of American adults own a Smartphone*. Pew Research Center. Retrieved June 17, 2013, from, http://www.pewinternet.org/~media/Files/Reports/2011/PIP_Smartphones.pdf

State-Specific Prevalence of Obesity Among Adults—United States, 2007. Morbidity and Mortality Weekly Report, Centers for Disease Control and Prevention. Retrieved June 17, 2013, from, <http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5728a1.htm>

9.2 Statistical Inference for Two Population Means with Unknown Population Standard Deviations

Baseball-Almanac. (2013). World Series History. In *Baseball-Almanac, 2013*. Retrieved June 17, 2013, from, <http://www.baseball-almanac.com/ws/wsmenu.shtml>

Data. (n.d.). Graduating Engineer + Computer Careers. <http://www.graduatingengineer.com>

Data. (n.d.). In *Microsoft Bookshelf*.

Data. (n.d.). United States Senate. Retrieved June 17, 2013, from <https://www.senate.gov/>

List of current United States Senators by Age. (n.d.). In *Wikipedia*. http://en.wikipedia.org/wiki/List_of_current_United_States_Senators_by_age

Sectoring by Industry Groups. (n.d.). Nasdaq. Retrieved June 17, 2013, from, <http://www.nasdaq.com/markets/barchart-sectors.aspx?page=sectors&base=industry>

Strip Clubs: Where Prostitution and Trafficking Happen. (2013). Prostitution Research & Education. Retrieved June 17, 2013, from <https://prostitutionresearch.com/strip-clubs-where-prostitution-and-trafficking-happen/>

9.4 Statistical Inference for Two Population Proportions

Data. (n.d.). American Cancer Society. Retrieved June 17, 2013, from, <http://www.cancer.org/index>

Data. (1994, November). Chancellor's Office, California Community Colleges.

Data. (December). *Educational Resources*.

Data. (n.d.). Hilton Hotels. Retrieved June 17, 2013, from, <http://www.hilton.com>

Data. (n.d.). Hyatt Hotels. Retrieved June 17, 2013, from, <http://hyatt.com>

Data. (n.d.). Statistics. United States Department of Health and Human Services.

Data. (n.d.). Whitney Exhibit on loan to San Jose Museum of Art.

State of the States. (2013). Gallup. Retrieved June 17, 2013, from, <http://www.gallup.com/poll/125066/State-States.aspx?ref=interactive>

West Nile Virus. Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), Division of Vector-Borne Diseases (DVBD). Retrieved June 17, 2013, from, <http://www.cdc.gov/ncidod/dvbid/westnile/index.htm>

10.2 Statistical Inference for a Single Population Variance

AppleInsider Price Guides. (n.d.). Apple Insider. Retrieved June 17, 2013, from, http://appleinsider.com/mac_price_guide

Data. (n.d.). World Bank.

10.3 The Goodness-of-Fit Test

Current Population Reports. (n.d.). U.S. Census Bureau.

Data. (n.d.). The College Board. <http://www.collegeboard.com>.

Data. (n.d.). U.S. Census Bureau.

Ma, Y., Bertone, E. R., Stanek III, E. J., Reed, G. W., Hebert, J. R., Cohen, N. L., Merriam, P. A., & Ockene, I. S. (2003, July 1). Association between Eating Patterns and Obesity in a Free-living US Adult Population. *American Journal of Epidemiology*, 158(1), 85-92. <https://doi.org/10.1093/aje/kwg117>

Ogden, C. L., Carroll, M. D., Kit, B. K., & Flegal, K. M. (2012, January). *Prevalence of Obesity in the United States, 2009–2010. NCHS Data Brief no. 82*. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. Retrieved May 24, 2013 from <http://www.cdc.gov/nchs/data/databriefs/db82.pdf>

Stevens, B. J. (n.d.). *Multi-family and Commercial Solid Waste and Recycling Survey*. Arlington County. Retrieved May 24, 2013, from, <http://www.arlingtonva.us/departments/EnvironmentalServices/SW/file84429.pdf>

10.4 The Test of Independence

DiCamilo, M., & Field, M. (2013, February 14). *Most Californians See a Direct Linkage between Obesity and Sugary Sodas. Two in Three Voters Support Taxing Sugar-Sweetened Beverages If Proceeds are Tied to Improving School Nutrition and Physical Activity Programs*. The Field Poll. Retrieved May 24, 2013, from, <http://field.com/fieldpollonline/subscribers/Rls2436.pdf>

Favorite Flavor of Ice Cream. (2016, October 22). Statistic Brain Research Institute. <http://www.statisticbrain.com/favorite-flavor-of-ice-cream>

Youngest Online Entrepreneurs List. (2016, June 29). Statistic Brain Research Institute. <http://www.statisticbrain.com/youngest-online-entrepreneur-list>

11.2 Statistical Inference for Two Population Variances

MLB Vs. Division Standings – 2012. (n.d.). ESPN. http://espn.go.com/mlb/standings/_/year/2012/type/vs-division/order/true

11.3 One-Way ANOVA and Hypothesis Tests for Three or More Population Means

Data. (n.d.). Fourth-grade classroom in 1994 in a private K – 12 school, San Jose, CA.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., & Ostrowski, E. (1994). *A Handbook of Small Datasets: Data for Fruitfly Fecundity*. Chapman & Hall.

MLB Standings – 2012. ESPN. http://espn.go.com/mlb/standings/_/year/2012

Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992). A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich. *Journal of the American Medical Association*, 268, 1578-1580.

12.1 Linear Equations

Data. (n.d.). Centers for Disease Control and Prevention.

Data. (n.d.). National Center for Agency Reporting Flu Cases and TB Prevention.

12.4 The Regression Equation

Data. (n.d.). Centers for Disease Control and Prevention.

Data. (n.d.). National Center for Agency Reporting Flu Cases and TB Prevention.

Data. (n.d.). National Center for Health Statistics.

Data. (n.d.). United States Census Bureau. http://www.census.gov/compendia/statab/cats/transportation/motor_vehicle_accidents_and_fatalities.html

VERSIONING HISTORY

This page provides a record of edits and changes made to this book since its initial publication. Whenever edits or updates are made in the text, we provide a record and description of those changes here. If the change is minor, the version number increases by 0.1. If the edits involve a number of changes, the version number increases to the next full number.

The files posted alongside this book always reflect the most recent version.

Version	Date	Change	Affected Web Page
1.0	March 2025	First Publication	N/A